National Center for Education Statistics

National Assessment of Educational Progress

1994 Trial State Assessment Program in Reading Secondary-Use Data Files User Guide

AUGUST 1996

Patricia E. O'Reilly
Christine A. Zelenak
Alfred M. Rogers
Debra L. Kline

Prepared by Educational Testing Service under contract with the National Center for Education Statistics
Office of Educational Research and Improvement ■ U.S. Department of Education

TABLE OF CONTENTS

Chapter 1 Introduction 1

1.1 What is NAEP? 1
1.2 Overview of the 1994 NAEP Trial State Assessment 1
1.2.1 Special Considerations 3
1.3 The NAEP Secondary-Use Trial State Assessment Data Files 3
1.4 Item Security 4
1.5 How to Use this Guide 5
1.6 An Analysis Example Using 1994 NAEP Data 6
1.6.1 Beginning the Analysis 6
1.6.2 Completing the Analysis with SPSS or SAS 8
1.6.3 Completing the Analysis with Statistical or Procedural Languages Other than SPSS or SAS 11
1.6.4 Error Estimation 11

Chapter 2 Special Considerations for Users 13

2.1 Introduction 13
2.2 The National Comparison Sample of Students 13
2.3 Partially Balanced Incomplete Block (PBIB) Spiral Method of Administration 13
2.4 Reporting Subgroups and Other Variables 13
2.5 Response Data from Teachers 14
2.6 Using Weights 14
2.7 Error Estimation 14
2.8 Monitored and Unmonitored Assessment Sessions 14
2.9 Revisions for the 1992 and 1994 Trial State Assessment Reading Data 15

Chapter 3 Instrument Design 17

3.1 Introduction 17
3.2 Student Assessment Booklets 17
3.2.1 Booklet Content 17
3.2.2 Booklet Assembly 21
3.2.3 Release Status for Item Blocks 21

3.3 Questionnaires 23

3.3.1 Student Questionnaires 23
3.3.2 IEP/LEP Student Questionnaire 23
3.3.3 Teacher Questionnaire 24
3.3.4 School Questionnaire 24

Chapter 4 Sample Selection and Weights 25

4.1 Introduction 25
4.2 Sample Selection 25

4.2.1 Selection of Schools 27

4.2.1.1 Frame Construction 27
4.2.1.2 Stratification 27
4.2.1.3 Selection of School Sample 29

4.2.2 Selection of Student Samples 30

4.3 Weighting Procedures 31

4.3.1 Full-Sample Weights 31
4.3.2 Replicate Weights 32
4.3.3 Summary of Weights and Their Use 32

Chapter 5 Data Collection, Materials Processing, Professional Scoring, and Database Creation 35

5.1 Introduction 35
5.2 Data Collection and Field Administration 35
5.3 Materials Processing and Data Entry 35
5.4 Professional Scoring of Reading Items 36

5.4.1 Description of Scoring 36
5.4.2 Constructed-Response Scores in the Secondary-Use Data Files 39

5.5 Database Creation 39

Chapter 6 Reporting Subgroups and Other Variables 41

6.1 Introduction 41
6.2 Reporting Subgroups for the 1994 Trial State Assessment 41
6.3 Variables Derived from the Student and Teacher Questionnaires 44

6.4 Variables Derived from Cognitive Items 46
6.5 Variables Related to Proficiency Scaling 47
6.6 Quality Education Data Variables (QED) 48

Chapter 7 NAEP Scaling Procedures and Their Application in the Trial State Assessment 49

7.1 Overview 49

7.2 Background 49
7.3 Scaling Methodology 50

7.3.1 The Scaling Models 50
7.3.2 An Overview of Plausible Values Methodology 53
7.3.3 Computing Plausible Values in IRT-Based Scales 54

7.4 NAGB Achievement Levels 55
7.5 Analyses 56

7.5.1 Computational Procedures 56
7.5.2 Statistical Tests 57
7.5.3 Biases in Secondary Analyses 57

7.6 Scaling the 1994 Trial State Assessment Reading Data 58

7.6.1 Item Response Theory (IRT) Scaling 59
7.6.2 Item Parameter Estimation 59
7.6.3 Estimation of State and Subgroup Proficiency Distributions 60
7.6.4 Linking State and National Scales 62
7.6.5 Producing a Reading Composite Scale 64
7.6.6 Proficiency Means for the 1994 Trial State Assessment Reading Scales . . 64

Chapter 8 Conducting Statistical Analyses with NAEP Data 67

8.1 Introduction 67
8.2 Using Weights to Account for Differential Representation 68

8.2.1 The 1994 State Samples of Students 68
8.2.2 Weights for Comparing Monitored and Unmonitored Sessions 69
8.2.3 The Comparison Sample from the National Assessment 70
8.2.4 School-Based Weights 70

8.3 Procedures Used by NAEP to Estimate Sampling Variability (Jackknifing) 70

8.3.1 Degrees of Freedom of the Jackknife Variance Estimate 74
8.3.2 Estimation of Subpopulations with Appropriate Jackknife Standard Errors 75

8.4 Procedures Used by NAEP to Handle Imprecision of Individual Measurement 75
8.5 Approximations 77

8.5.1 Approximations for Sampling Variability 77
8.5.2 Approximations for Measurement Error Variability 79
8.5.3 Approximating Both Sampling and Measurement Variability 79

8.6 Additional Sources of Error 80
8.7 A Note Concerning Multiple Comparisons 80

Chapter 9 Content and Format of Data Files, Layouts, and Codebooks 83

9.1 Introduction 83
9.2 Data Files 83

9.2.1 Respondent Data 83
9.2.2 SPSS and SAS Control Statement Files 86
9.2.3 Machine-Readable Catalog Files 86

9.3 Printed Documentation 90

9.3.1 File Layouts 90
9.3.2 Codebooks 91

Chapter 10 Working with SPSS and SAS 93

10.1 Introduction 93
10.2 SPSS and SAS Control Statement Files 93
10.3 Creating SPSS System Files 94
10.4 Creating SAS System Files 94
10.5 Merging Files Under SPSS or SAS 97
10.6 Computing the Estimated Variance of a Mean (Jackknifing) Using SPSS or SAS . . 99
10.7 An Analysis Example Using NAEP Data with SPSS and SAS 103

Appendix A NAEP History 111

Appendix B IRT Parameters for Reading Items 115

Appendix C Glossary of Terms 125

Appendix D Unweighted Nonpublic-School Data Files 129

References Cited in Text 131

LIST OF TABLES & FIGURES

Table 1-1 Jurisdictions Participating in the 1994 Trial State Assessment Program 2
1-2 SAS Analysis Example Output 7
1-3 SPSS Analysis Example Output 7
1-4 NAEP Variables Used to Produce the Analysis 8
1-5 SAS Code to Produce Example Analysis 9
1-6 SPSS Code to Produce Example Analysis 10

3-1 Percentage Distribution of Items by Grade and Reading Purpose 20
3-2 Percentage Distribution of Items by Reading Stance for Grades 4, 8, and 12 20
3-3 Cognitive and Noncognitive Block Information 22
3-4 Booklet Contents 22

4-1 1994 Trial State Assessment Participation 26
4-2 Summary of Weights for the 1994 Trial State Assessment 33

5-1 1994 NAEP Trial State Assessment Number of Constructed-Response Items in Each Range of Percentages of Exact Agreement Between Readers 37

6-1 NAEP Geographic Regions 45
6-2 Scaling Variables for the 1994 Trial State Assessment Sample 48

7-1 Extended Constructed-Response Items, 1994 Trial State Assessment in Reading 61
7-2 Transformation Constants for the 1994 Trial State Assessment 64
7-3 Weights Used for Each Scale to Form the Reading Composite 65
7-4 Average Reading Proficiencies by Scale and Plausible Value 1994 National Reading Grade 4 Public-School Comparison 65

8-1 Example Dataset to Demonstrate the Jackknife 74

9-1 NAEP 1994 State Reading Data Package Description: Grade 4 84
9-2 NAEP 1994 Reading File Record Counts by Jurisdiction 85
9-3 Special Response Codes 86
9-4 NAEP 1994 State Machine-Readable Catalog File Layout 87
9-5 Scaling Categories and Codes 90

10-1 SPSS Control Statement Synopsis 95
10-2 SAS Control Statement Synopsis 96
10-3 Matching School and Student Files 98
10-4 Matching School and Excluded Student Files, State Sample 98
10-5 Standard Error Computation: Multiweight Method Using SPSS 99
10-6 Standard Error Computation: Multiweight Method Using SAS 100
10-7 Standard Error Computation: Multiweight Method Using SPSS with Correction for Imputation 101
10-8 Standard Error Computation: Multiweight Method Using SAS with Correction for Imputation . 102

Table 10-9 SPSS Analysis Example Using Jackknife Standard Error Estimates 103
10-10 SAS Analysis Example Using Jackknife Standard Error Estimates 104
10-11 NAEP Variables Used to Produce the Analysis 105
10-12 SPSS Code for Steps 2 through 7 to Produce Example Analysis 106
10-13 SAS Code for Steps 2 through 7 to Produce Example Analysis 108

A-1 National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1994 112

B-1 IRT Parameters for the Trial State Assessment, Reading for Literary Experience Scale, Grade 4 116
B-2 IRT Parameters for the Trial State Assessment, Reading to Gain Information Scale, Grade 4 118
B-3 IRT Parameters for the National Reading Samples, Reading for Literary Experience Scale Age 9/Grade 4 120
B-4 IRT Parameters for the National Reading Samples, Reading to Gain Information Scale Age 9/Grade 4 122

D-1 Special Data File Record Counts by State 130

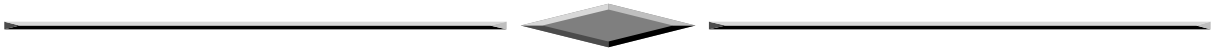

Figure 3-1 Description of Reading Stances 18
3-2 Description of Purposes for Reading 19

5-1 Extended Constructed-Response Scoring Guide for “Sybil Sounds the Alarm” 38

ACKNOWLEDGEMENTS

The NAEP 1994 Trial State reading data are a unique and valuable source of information for the research community. To extend the usefulness of these data, we have provided the secondary-use data files and user guide. We hope the files will be used by many researchers for secondary data analysis.

The data files were created through the efforts of a talented and dedicated team of data analysts. Special acknowledgment must be given to Alfred Rogers, who developed the sophisticated database systems, and Patricia O'Reilly, who created the secondary-use data files. Kate Pashley and David Freund were responsible for the creation of the database. Special thanks go to Steven Isham, who performed the data analyses with the assistance of Lois Worthington. State report design and development were led by Laura Jerry; the data compendium was produced by Jennifer Nelson. Others collaborating on the reports were Phillip Leung, Bruce Kaplan, Inge Novatkoski, Steven Isham, and Dave Freund. Additional analysis support was provided by Laura Jenkins, Michael Narcowich, and Minhwei Wang.

The user guide has been particularly enhanced through the contributions of Statistical and Psychometric Research staff members Nancy Allen, Eugene Johnson, and Robert Mislevy. We are also grateful to Keith Rust for his contributions and comments.

The user guide was designed and produced under the outstanding editorial supervision of Christine Zelenak and Debra Kline.

John L. Barone
Director of Data Analysis
National Assessment of
Educational Progress

CHAPTER 1

INTRODUCTION

1.1 What is NAEP?

The National Assessment of Educational Progress (NAEP) is an ongoing, congressionally mandated national survey of the knowledge, skills, understanding, and attitudes of young Americans in major subjects usually taught in school. Its primary goals are to detect and report the status of and long-term changes in the educational attainments of young Americans. The purpose of NAEP is to gather information that will aid educators, legislators, and others in improving the educational experience of youth in the United States. It is the first ongoing effort to obtain comprehensive and dependable national achievement data in a uniform, scientific manner.

NAEP began in 1969 as an annual survey of American students ages 9, 13, and 17 in various subject areas; young adults ages 26 to 35 were surveyed less frequently. Since the 1980-81 school year, budget restraints have prompted a shift to biennial data collection. In the 1984 assessment, NAEP began sampling students by grade as well as age.

The 1994 Trial State Assessment Program once again assessed the reading skills and understanding of representative samples of fourth-grade students in participating jurisdictions. The participation of jurisdictions in the Trial State Assessment has been, and continues to be, voluntary. The 1994 program broke new ground in two ways. The 1994 NAEP authorization called for the assessment of samples of both public and private school students. Thus, for the first time in NAEP, jurisdiction-level samples of students from Catholic schools, other religious schools and private schools, Domestic Department of Defense Education Activity schools, and Bureau of Indian Affairs schools were added to the Trial State program. Second, samples of students from the Department of Defense Education Activity overseas schools participated as a jurisdiction, along with the states and territories that have traditionally had the opportunity to participate in Trial State Assessment Program.

In April 1988, Congress reauthorized NAEP and added a new dimension to the program—voluntary state-by-state assessments on a trial basis in 1990 and 1992, in addition to continuing the national assessments that NAEP has conducted since its inception.

More information about NAEP and its history is provided in Appendix A.

1.2 Overview of the 1994 NAEP Trial State Assessment

The first NAEP Trial State Assessment was conducted in 1990. The program collected information on the mathematics knowledge, skills, understanding, and perceptions of a representative sample of eighth-grade students in public schools in 37 states, the District of Columbia, and two territories. The second phase of the Trial State Assessment Program, conducted in 1992, collected similar mathematics data for representative samples of fourth- and eighth-grade students and assessed the reading knowledge, skills, understanding, and perceptions of a representative sample of fourth-grade students in public schools in 41 states, the District of Columbia, and two territories. The third NAEP Trial State Assessment once again assessed the reading skills and understanding of representative samples of fourth-grade students in participating jurisdictions.

Table 1-1 lists the jurisdictions that participated in the 1994 Trial State Assessment Program. More than 120,000 fourth-grade students participated in the reading assessments in those jurisdictions. The students were administered the same reading assessment booklets that were used in NAEP's 1994 national fourth-grade reading assessment.

The reading framework that guided both the 1994 Trial State Assessment and the 1994 national assessment is the same framework used for the 1992 NAEP assessment. The framework was developed for NAEP through a consensus project of the Council of Chief State School Officers, funded by the National

Table 1-1
Jurisdictions Participating in the 1994 Trial State Assessment Program

JURISDICTIONS			
Alabama	Guam	Minnesota	Pennsylvania
Arizona	Hawaii	Mississippi	Rhode Island
Arkansas	Idaho	Missouri	South Carolina
California	Indiana	Montana*	Tennessee
Colorado	Iowa	Nebraska	Texas
Connecticut	Kentucky	New Hampshire	Utah
Delaware	Louisiana	New Jersey	Virginia
DoDEA Overseas*	Maine	New Mexico	Washington*
District of Columbia**	Maryland	New York	West Virginia
Florida	Massachusetts	North Carolina	Wisconsin
Georgia	Michigan	North Dakota	Wyoming

**Note: Washington, Montana, and DoDEA (Department of Defense Education Activity) overseas schools participated in the 1994 program but did not participate in the 1992 program.*
***Note: The District of Columbia participated in the testing portion of the 1994 Trial State Assessment Program. However, in accordance with the legislation providing for participants to review and give permission for release of their results, the District of Columbia chose not to publish their results in the reports.*

Assessment Governing Board. Hence, 1994 provides the first opportunity to report jurisdiction-level trend data for a NAEP reading instrument for those states and territories that participated in both the 1992 and 1994 Trial State Assessment programs. In addition, questionnaires completed by the students, their reading teachers, and principals or other school administrators provided an abundance of contextual data within which to interpret the reading results.

Educational Testing Service (ETS) was the contractor for the 1994 NAEP programs, including the Trial State Assessment. ETS was responsible for overall management of the programs and development of the overall design, the items and questionnaires, data analysis, and reporting. Westat, Inc. was responsible for all aspects of sampling and field operations. National Computer Systems (NCS) was responsible for the printing, distribution, and receipt of assessment materials; the scanning of assessment data; and the professional scoring of constructed responses.

These secondary-use files contain the data that were used to create a series of reports that have been prepared for the 1994 Trial State Assessment Program in reading, including:

- A *State Report* for each participating jurisdiction that describes the reading proficiency of the fourth-grade public- and nonpublic-school students in that jurisdiction and relates their proficiency to contextual information about reading policies and instruction.

- The report *1994 NAEP Reading: A First Look*, which provides overall public-school results and results for major NAEP reporting subgroups for all jurisdictions that participated in the Trial State Assessment Program, as well as selected results from the 1994 national reading assessment.

- The *NAEP 1994 Reading Report Card for the Nation and the States*, which provides both public- and nonpublic-school data for all jurisdictions that participated in the Trial State Assessment Program

along with a more complete report of the results from the 1994 national reading assessment.

- *Results from the NAEP 1994 Reading Assessment: At A Glance*, providing the highlights of the *Reading Report Card*.

- The *Cross-State Data Compendium from the NAEP 1994 Reading Assessment*, which includes jurisdiction-level results for all demographic, instructional, and experiential background variables included in the *Reading Report Card* and *State Report*.

- *Data Almanacs* for each jurisdiction that contain a detailed breakdown of the reading proficiency data according to the responses to the student, teacher, and school questionnaires for the public-school, nonpublic-school, and combined populations as a whole and for important subgroups of the public-school population. There are six sections to each almanac:

  - The *Distribution Data Section* provides information about the percentages of students at or above the three composite-scale achievement levels (and below basic). For the composite scale and each reading scale (Reading for Literary Experience and Reading to Gain Information), this almanac also provides selected percentiles for the public-school, nonpublic-school, and combined populations and for the standard demographic subgroups of the public-school population.

  - The *Student Questionnaire Section* provides a breakdown of the composite-scale proficiency data according to the students' responses to questions in the three student questionnaires included in the assessment booklets.

  - The *Teacher Questionnaire Section* provides a breakdown of the composite-scale proficiency data according to the teachers' responses to questions in the reading teacher questionnaire.

  - The *School Questionnaire Section* provides a breakdown of composite-scale proficiency data according to the principals' (or other administrators') responses to questions in the school characteristics and policies questionnaire.

  - The *Scale Section* provides a breakdown of the proficiency data for the two purpose-for-reading scales according to selected items from the questionnaires.

  - The *Reading Item Section* provides the response data for each reading item in the assessment.

1.2.1 Special Considerations

Because of the complexity of the NAEP design (see Chapters 3 and 4), data file users need some understanding of the design before performing analyses. Special characteristics of the assessment are outlined in Chapter 2.

The data files contain sampling weights for each student that should be used in statistical analyses. In addition, because of the complex sampling scheme, conventional methods of standard error estimation do not produce appropriate estimates. The NAEP sampling design also reduces the effective degrees of freedom for statistical analysis. These issues are discussed in Chapter 8.

1.3 The NAEP Secondary-Use Trial State Assessment Data Files

Prior to 1990, a "public-use" version of the NAEP data files was distributed to secondary users. However, in order to comply with 5 U.S.C. 552a and U.S.C. 1221e-1, only a "restricted-use" version of the 1994 NAEP data files will be distributed for secondary use (this procedure was also followed for the 1990 and 1992 data files). These will be loaned to states and people designated by them under a licensure procedure designed to assure confidentiality of identifiable district, school, and individual data.

The secondary-use files for each state contain data for students, teachers, schools, and excluded students in the state and for students, teachers, and schools in the sample from the national reading assessment that was used for comparisons between the nation and the state. The April 1996 version of the files represents the first release of these data. The secondary-use data files contain:

- students' responses to cognitive reading items;
- students' responses to questions about their demographic backgrounds and educational experiences;
- information about students' schools and reading teachers;
- information about students excluded from the assessment (state sample only);
- sampling weights for students, schools, and (for state sample only) excluded students;
- proficiency scale scores for the reading composite scale and two purpose-for-reading scales—Reading to Gain Information and Reading for Literary Experience.
- machine-readable catalog files; and
- SPSS and SAS control statement files.

The data files can be used in a variety of computing environments and are available in the following forms:

- *CD-ROM disk*: ASCII (uncompressed) format;
- *9-track tape reel (6,250 bpi)*: blocked EBCDIC (uncompressed) format; and
- *IBM 3480 tape cartridge (38,000 bpi)*: blocked EBCDIC (compressed/uncompressed) format.

To use the files, you will need an IBM PC-compatible workstation with a CD-ROM drive or a mini- or mainframe computer with the appropriate tape drive.

Codebooks for each state provide the layout of the data, a description of each variable, and a description of each raw data file for both the state and the sample from the national reading assessment that was used for comparisons between the nation and the state. The content and format of the data files and codebooks are described in Chapter 9. Table 9-1 in that chapter gives the files for each sample and the record lengths for each file.

If you have questions about the data files and their use, contact:

Mr. Robert Clemons
National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
Education Assessment Division, Room 308F
555 New Jersey Avenue, N.W.
Washington, DC 20208-5653
(202) 219-1690 or Bob_Clemons@ed.gov

1.4 Item Security

In accordance with federal legislation regarding security of NAEP items and guidelines designed by the National Center for Education Statistics, each NAEP cognitive item has been assigned a release status. *Public release* items are available for unrestricted public use. *Secured release* items are available only to users who have agreed to conditions designed to ensure item security and to prevent misuse of items. Items not classified as either public or secured release are reserved exclusively for NAEP use—for example, for administration in future assessments to allow analysis of trends in performance levels. To preserve the integrity of NAEP, it is essential that these items remain secure.

The data files and codebooks contain response counts for all items used in the assessment and a short descriptive label for each item. For each cognitive reading item that has not been classified as either public or secured release, text describing response options (for multiple-choice items) or scoring categories (for constructed-response items) has been replaced with generic descriptions.

All student demographic and reading background items and items from teacher, school, and excluded student questionnaires are classified as public release and are available to secondary users.

1.5 How to Use This Guide

Chapters 2 through 10 and the appendices provide detailed information about the 1994 Trial State Assessment, the data files, and recommended methods of working with the data to perform analyses. A summary of these chapters follows.

Chapter 2: Special Considerations for Users

This chapter describes features of the assessment design and assessment data that may be of special concern to researchers who wish to perform their own analyses of the data.

Chapter 3: Instrument Design

This chapter includes a description of the content, organization, and method of administration for the student assessment booklets and the teacher, excluded student, and school questionnaires.

Chapter 4: Sample Selection and Weights

This chapter explains the methods by which schools, students, and teachers were chosen to be included in the assessment; the method by which some students were chosen for the sample but subsequently excluded from the assessment; and the sampling weights included on the data files.

Chapter 5: Data Collection, Materials Processing, Professional Scoring, and Database Creation

Assessment administration, data entry and editing, scoring of constructed-response items, and creation of the NAEP database are all described in this chapter.

Chapter 6: Reporting Subgroups and Other Variables

This chapter describes the NAEP reporting subgroups, derived and composite variables from the background questionnaires, composite variables created for the NAEP reports, item response theory (IRT) variables, and other data variables that are not self-explanatory.

Chapter 7: NAEP Scaling Procedures and Their Application in the Trial State Assessment

This chapter provides an overview of the scaling methodologies used by NAEP, the scale-score analyses carried out in the 1994 Trial State Assessment, and supporting information on the scale-score variables that appear on the data files.

Chapter 8: Conducting Statistical Analyses with NAEP Data

This chapter discusses the weights on the data files, how to use them in different types of analyses, and methods for estimating sampling variability and measurement error.

Chapter 9: Content and Format of Data Files, Layouts, and Codebooks

Detailed descriptions of the raw data files, layouts, codebooks, machine-readable catalogs, and SPSS and SAS control statement files are found in this chapter.

Chapter 10: Working with SPSS and SAS

This chapter provides procedures for creating SPSS and SAS system files, merging data files, and using the jackknife procedure to estimate standard errors, as well as an example of how to analyze NAEP data with SPSS and SAS.

Appendix A provides information about the history of NAEP.

Appendix B contains IRT parameters for each cognitive item used in the scaling of the reading data.

Appendix C is a glossary of terms.

Appendix D contains unweighted nonpublic-school data files.

References Cited in Text provide complete information on sources cited in the text.

1.6 An Analysis Example Using 1994 NAEP Data

This section presents an example of how to produce a simple descriptive analysis table from the national comparison sample data files that are used for state/nation comparisons. The example could be carried out in a similar way for each state's files. Most analyses of NAEP data can be performed in four basic steps:

- Identify and access the appropriate data file
- Identify and extract the relevant variables
- Select the proper subset of students
- Compute and print the results

The method you choose to perform these steps may vary with the complexity of the analysis or with the statistical or procedural language you are using.

To aid users, we have included three types of files:

- machine-readable catalog files
- SAS control statement files
- SPSS control statement files

The machine-readable catalog files can be used with any statistical or procedural language to quickly extract and store the location and labeling information for every field on the NAEP data files. This information can then be used by your program to extract actual response data from the data files. There is a catalog file for each data file; each catalog file contains a record for every field in the corresponding data file (more about the machine-readable catalog files can be found in Chapter 9).

For SPSS and SAS users, control statement files are provided to facilitate the creation of SPSS and SAS system files. There are SPSS and SAS control files for each data file. Part of each control file contains the field name, location, and format for each variable on the corresponding data file (more about control statement files can be found in Chapter 10).

1.6.1 Beginning the Analysis

The analysis in our example produced the following estimates of the mean reading proficiency level for fourth-grade public-school girls in the national comparison sample by the amount of television watched each day. The output from SAS is given in Table 1-2; the output from SPSS is shown in Table 1-3.

To begin this analysis, you need to identify

- the file that contains response data for the national comparison sample of fourth-grade students and
- the relevant variables in the file.

NAEP files are described in Chapter 9 and listed in Table 9-1; the correct file for our example is 'NCR1STUD.DAT'. Next, find the data set record layout for 'NCR1STUD.DAT' in the accompanying codebook. Here you will find the names and file locations of the variables needed to produce this table (unweighted response counts for each variable are found in the corresponding codebook). Five variables (described in Table 1-4) are required to produce the analysis: SCHTYPE, DSEX, ORIGWT, B001801A, and RRPCM1.

Because this example is relatively simple (requiring the use of only five variables), you can manually enter the variable labels and locations into your computer program. For analyses that require many variables, you should use the machine-readable catalog files or, if you are a SPSS or SAS user, the control statement files.

Section 1.6.2 describes how to complete the analysis using the statistical packages SPSS and SAS. Section 1.6.3 describes how to use the machine-readable catalog files to complete the analysis using statistical or procedural languages other than SPSS or SAS. In Section 1.6.4, we discuss the importance of the proper estimation of standard errors.

Table 1-2
SAS Analysis Example Output

1994 National Comparison Sample
Reading Results for 4th Grade Public-School Girls
by Amount of Television Viewing

OBS	HOW MUCH TELEVISION DO YOU USUALLY WATCH	WEIGHTED N	PERCENT	MEAN
1	NONE	24226.55	1.5788	214.883
2	1 HOUR OR LESS	300946.57	19.6120	220.316
3	2 HOURS	343473.39	22.3834	224.486
4	3 HOURS	265853.73	17.3251	224.095
5	4 HOURS	199259.39	12.9853	224.773
6	5 HOURS	126470.41	8.2418	212.321
7	6 HOURS OR MORE	274272.98	17.8737	197.000

Table 1-3
SPSS Analysis Example Output

1994 National Comparison Sample
Reading Results for 4th Grade Public-School Girls
by Amount of Television Viewing

HOW MUCH TELEVISION DO YOU USUALLY WATCH	WEIGHTED N	PERCENT	MEAN
NONE	24226.55	1.579	214.883
1 HOUR OR LESS	300946.57	19.612	220.316
2 HOURS	343473.39	22.383	224.486
3 HOURS	265853.73	17.325	224.095
4 HOURS	199259.39	12.985	224.773
5 HOURS	126470.41	8.242	212.321
6 HOURS OR MORE	274272.98	17.874	197.000

Table 1-4
NAEP Variables Used to Produce the Analysis

Seq. No.	Field Name	Column Position	Field Width	Decimal Places	Type	Range	Short Label
28	SCHTYPE	68	1	–	D	1-5	School type
36	DSEX	94	1	–	D	1-2	Gender
50	ORIGWT	175	7	2	C	–	Student weight (unadjusted)
213	RRPCM1	896	5	2	C	–	Plausible NAEP reading value #1 (Composite)
229	B001801A	932	1	–	D	1-7	How much television do you usually watch each day?

1.6.2 Completing the Analysis with SPSS or SAS

You can use any statistical computing language or package to access the raw data file, extract the relevant variables, select the proper subset of students, and compute the table. In this section, we carry out the rest of the analysis using the statistical packages SPSS and SAS.

1) Select the file containing the fourth-grade students in the national comparison sample. This is one of the samples described in Table 9-1 in Chapter 9; its file name is NCR1STUD.DAT. Identify the relevant variables from the data set record layout: SCHTYPE, DSEX, ORIGWT, RRPCM1, and B001801A.

2) From the raw data file NCR1STUD.DAT select the appropriate subset of students for the table. This selection restricts the analysis to public-school (SCHTYPE=1) girls (DSEX=2) who have valid reading proficiency (RRPCM1) and television viewing (B001801A) values. This analysis will be weighted to the population using ORIGWT as the weighting factor.

3) Compute overall weighted counts for use in the computation of percentages.

4) Compute weighted counts and sums for each level of television viewing (B001801A).

5) Merge the aggregates from steps 3 and 4 and compute percentages and means.

6) Print the final result in a formatted table.

The SAS code for performing the analysis is shown in Table 1-5; the SPSS code for the analysis is shown in Table 1-6.

Please note that this example does not include standard error estimates that account for NAEP sampling design and measurement error components. In Chapter 10, we provide a second version of this example that demonstrates the proper computation of standard error estimates.

Table 1-5

SAS Code to Produce Example Analysis

```
TITLE1 '1994 National Comparison Sample';
TITLE2 'Reading Results for 4th Grade Public-School Girls';
TITLE3 'by Amount of Television Viewing';
/***************************** STEP 1 ******************************/
DATA    A;
INFILE   'G:\DATA\NCR1STUD.DAT'  LRECL=1524;
INPUT
  SCHTYPE      68      DSEX             94       ORIGWT      175-181 .2
  B001801A    932      RRPCM1      896-900 .2 ;
/***************************** STEP 2 ******************************/
IF (RRPCM1   NE .);
IF (DSEX     EQ 2);
IF (SCHTYPE  EQ 1);
IF (B001801A NE .) AND
   (B001801A GT 0) AND
   (B001801A LT 8);
WTX = ORIGWT*RRPCM1;
MDUMMY = 0;
KEEP DSEX ORIGWT B001801A RRPCM1 WTX MDUMMY;
LABEL
  DSEX     = 'GENDER                                        '
  ORIGWT   = 'STUDENT WEIGHT (UNADJUSTED)                   '
  B001801A = 'HOW MUCH TELEVISION DO YOU USUALLY WATCH'
  RRPCM1   = 'PLAUSIBLE NAEP READING VALUE #1 (COMP.) ';
PROC FORMAT;
  VALUE B001801A  .='TOTAL               '    1='NONE               '
                 2='1 HOUR OR LESS      '    3='2 HOURS            '
                 4='3 HOURS             '    5='4 HOURS            '
                 6='5 HOURS             '    7='6 HOURS OR MORE    ';
/***************************** STEP 3 ******************************/
PROC SUMMARY;
    VAR MDUMMY ORIGWT;
    OUTPUT OUT=B  SUM(MDUMMY ORIGWT) = MDUMMY TOTSWT;
/***************************** STEP 4 ******************************/
PROC SUMMARY DATA=A;
    CLASS B001801A;
    VAR ORIGWT WTX MDUMMY;
    OUTPUT OUT=C
        SUM(MDUMMY ORIGWT WTX) = MDUMMY SWT SWX;
/***************************** STEP 5 ******************************/
DATA D;
    MERGE B C;
    BY MDUMMY;
    IF (B001801A NE .);
    PCT = 100 * SWT/TOTSWT;
    XBAR=SWX/SWT;
/***************************** STEP 6 ******************************/
PROC PRINT SPLIT='*';
    FORMAT B001801A B001801A.;
    LABEL SWT  = 'WEIGHTED N'
          PCT  = 'PERCENT'
          XBAR = 'MEAN';
    VAR  B001801A SWT PCT XBAR;
RUN;
```

Table 1-6

SPSS Code to Produce Example Analysis

```
TITLE      "1994 National Comparison Sample:  Reading Results for".
SUBTITLE   "4th Grade Public-School Girls by Amount of TV Viewing".
FILE HANDLE  NCR1STUD  /NAME='G:\DATA\NCR1STUD.DAT'  /LRECL=1524.
* ---------------------------- STEP 1 -------------------------------.
DATA LIST FILE=NCR1STUD/
  SCHTYPE     68      DSEX             94       ORIGWT      175-181 (2)
  B001801A   932      RRPCM1      896-900 (2).
* ---------------------------- STEP 2 -------------------------------.
SELECT IF (NOT SYSMIS(RRPCM1)).
SELECT IF DSEX = 2.
SELECT IF SCHTYPE = 1.
SELECT IF B001801A LT 8.
COMPUTE WTX = ORIGWT*RRPCM1.
VARIABLE LABELS
  DSEX       'GENDER                                        '
  ORIGWT     'OVERALL STUDENT FULL-SAMPLE WEIGHT            '
  B001801A   'HOW MUCH TELEVISION DO YOU USUALLY WATCH'
  RRPCM1     'PLAUSIBLE NAEP READING VALUE #1 (COMP.) '.
VALUE LABELS
  B001801A                                    1 'NONE                '
                 2 '1 HOUR OR LESS      '    3 '2 HOURS             '
                 4 '3 HOURS             '    5 '4 HOURS             '
                 6 '5 HOURS             '    7 '6 HOURS OR MORE     '.
* ---------------------------- STEP 3 -------------------------------.
AGGREGATE   OUTFILE=TEMP1/ BREAK=DSEX/ TOTSW = SUM(ORIGWT).
* ---------------------------- STEP 4 -------------------------------.
AGGREGATE   OUTFILE=*/ BREAK=DSEX B001801A/
      SWT,SWX = SUM(ORIGWT WTX).
* ---------------------------- STEP 5 -------------------------------.
MATCH FILES  FILE=* / TABLE=TEMP1 / BY=DSEX.
COMPUTE  XBAR = SWX/SWT.
COMPUTE  PCT  = 100*(SWT/TOTSW).
PRINT FORMATS  SWT (F10.2)  PCT XBAR (F9.3).
* ---------------------------- STEP 6 -------------------------------.
REPORT
  / FORMAT = LIST AUTOMATIC ALIGN(CENTER) MARGINS(1,121)
  / TITLE = CENTER
       '1994 National Comparison Sample'
       'Reading Results for 4th Grade Public-School Girls'
       'by Amount of Television Viewing'
  / VARIABLES = B001801A (LABEL)  SWT 'WEIGHTED' 'N'
      PCT 'PERCENT'  XBAR 'MEAN'.
* ------------------------------------------------------------------.
```

1.6.3 Completing the Analysis with Statistical or Procedural Languages Other than SPSS or SAS

This section explains how to complete the sample analysis using the machine-readable catalog files. Each catalog file contains one record for every data field in its corresponding data file. These records describe the contents of each data field (e.g., field name, field location, response labels, range of data in the field, etc.). Table 9-4 in Chapter 9 contains a complete layout for the catalog files.

In our example, 'NCR1STUD.CAT' (see Table 9-1 in Chapter 9) is the machine-readable catalog file that corresponds to the student data file 'NCR1STUD.DAT'. Each record in this catalog file describes one of the fields in the student data file. To access the student data with the catalog file and complete the analysis:

1) Extract and store the field locations and labels for each variable required for analysis from the catalog file.

2) Using the stored information from the catalog file, read the student data file to extract and label the required student data fields.

3) In your program, perform the required analyses with the extracted variables (SCHTYPE, DSEX, ORIGWT, RRPCM1, and B001801A).

4) Print the results using the stored labeling information from the catalog file.

Please note that this procedure does not include standard error estimates that account for NAEP sampling design and measurement error components (see Section 1.6.4).

1.6.4 Error Estimation

The preceding example is presented as a practical introduction to the secondary-use data files. We have not attempted here to produce proper standard error estimates that account for NAEP sampling design and measurement error components. Such an accounting is required for statistical comparison of the results shown in our table. Because the NAEP sample is not a simple random sample, ordinary formulas for estimating the standard error of sample statistics will produce values that are too small.

Before attempting any analysis of NAEP data, users should understand the special characteristics of the NAEP sampling design (Chapters 2 and 4). Alternate methods for computing standard errors and recommended formulas for obtaining degrees of freedom are given in Chapter 8.

CHAPTER 2

SPECIAL CONSIDERATIONS FOR USERS

2.1 Introduction

Because of the complexity of the NAEP design, it is important for users to have some understanding of it before performing analyses of the data. The following sections highlight areas of potential importance to the user in conducting analyses.

Details of the design and data analysis for the 1994 Trial State Assessment are provided in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* (Mazzeo, Allen, & Kline, 1995).

2.2 The National Comparison Sample of Students

One of the purposes of the Trial State Assessment Program was to allow each participating jurisdiction to compare its results with those of the nation as a whole and with those of the geographic region in which that state is located.[1] To permit such comparisons, a nationally representative sample of fourth-grade students was assessed as part of the national assessment, using the same instruments that were used in the Trial State Assessment.

Because of differences between the state and national samples (described in Chapter 8), it was necessary to create a subsample from the full national sample to allow for valid state/nation comparisons. Data from this subsample (referred to as the national comparison sample) are included on the secondary-use data files, along with the appropriate weights to be used for analyses. Chapter 8 provides information on conducting analyses using the national comparison sample.

[1]No regions were designated for the territories.

2.3 Partially Balanced Incomplete Block (PBIB) Spiral Method of Administration

The term "partially balanced incomplete block (PBIB) spiral" refers to the method used to assemble assessment items into instruments. This method was developed to allow the study of the interrelationships among items within a subject area. As a result of this design, all items are given to approximately the same number of students, but no student receives all items.

The PBIB-spiral design for the reading booklets in the Trial State Assessment is discussed in Chapter 3.

2.4 Reporting Subgroups and Other Variables

In addition to reporting overall state or national achievement results, NAEP reports results for several student subgroups—gender, race/ethnicity, type of community, and level of parents' education. Some of these subgroups were derived from students' responses to one or more assessment items. Chapter 6 defines and explains the reporting subgroups.

Certain derived variables on the data files were created through the systematic combination of values from one or more items from the student, teacher, or school questionnaire. The derived variables are described in Chapter 6.

The files also contain reading proficiency variables, called plausible values. These variables, developed for scaling purposes, are described in Chapter 6; their explanation and use are given in Chapter 7.

Some variables on the files were taken from sources other than the assessment instruments. For optimal use of these variables, see their explanations in Chapter 6.

2.5 Response Data from Teachers

The reading teachers of the students assessed in both the national reading assessment and the Trial State Assessment were asked to complete a two-part questionnaire about their instructional practices, teaching backgrounds, and other characteristics. The first part of the questionnaire pertained to the teachers' background and training; the second pertained to the programs and instructional methods the teacher used for each class containing an assessed student.

In the NAEP data files, the data from the teacher questionnaire have already been linked with the appropriate student response data and included on the student data records, allowing correct and efficient analysis of the teacher/student data without requiring users to match data from separate files.

Note: The purpose of this sample is to estimate the numbers of students whose teachers have various attributes, not to estimate the attributes of the teacher population. Because of the nature of the sampling for the Trial State Assessment, the responses to the reading teacher questionnaire do not necessarily represent all fourth-grade reading teachers in a state. Rather, they represent the teachers of the particular students being assessed.

2.6 Using Weights

In the NAEP sampling design, students do not have an equal probability of being selected. Therefore, as in all such complex surveys, each student has been assigned a sampling weight. When computing descriptive statistics or conducting inferential procedures, one should weight the data properly for each student. Performing statistical analyses without weights can lead to misleading results.

Chapter 4 explains the weight variables and how they were developed; Chapter 8 explains how to use weights in performing analyses.

2.7 Error Estimation

The 1994 NAEP sampling design involved the selection of clusters of students from the same school, as well as clusters of schools from urbanicity, income, and minority strata (in the case of the Trial State Assessment) and from the same geographically defined primary sampling unit, or PSU (in the case of the national assessment). As a result, observations are not independent of one another as they are in a simple random sample. Therefore, use of ordinary formulas for estimating the standard error of sample statistics will result in values that are too small. Alternate methods of computing standard errors are provided in Chapter 8.

Another effect of the sampling design is a reduction of the effective degrees of freedom, which in the 1994 NAEP design are a function of the number of clusters of schools (for the Trial State Assessment) or clusters of PSUs (for the national assessment) and the number of strata in the design, rather than the number of subjects. Recommended formulas for obtaining degrees of freedom can be found in Chapter 8.

2.8 Monitored and Unmonitored Assessment Sessions

As part of the effort to ensure security and uniformity in the administration of the Trial State Assessment, random samples of the assessment sessions were monitored by trained quality control monitors. Within each state, and across all states, randomly equivalent samples of students received each block of cognitive items in a particular position within a booklet under monitored and unmonitored administration conditions. Thus, it was possible to conduct analyses comparing the data from the monitored sessions with the data from the unmonitored sessions. Details of the monitoring process are given in Section 8.2.2.

2.9 Revisions for the 1992 and 1994 Trial State Assessment Reading Data

In April 1995, results from the 1994 Trial State Assessment of reading were released as part of the report *1994 NAEP Reading: A First Look*. Subsequently, ETS/NAEP research scientists discovered an error in the documentation for the ETS version of the PARSCALE program, which was used to compute the 1994 NAEP scale score results. The error affected how omitted responses were treated in the IRT scaling of the extended constructed-response items that received partial-credit scoring. It was determined that the error had been introduced in the analysis of the 1992 NAEP data; hence, the 1992 state and national reading scales were also affected.

The analyses for 1992 and 1994 were subsequently redone; the *First Look* report was revised and reissued. The 1994 secondary-use data files contain the corrected results. A revised version of the secondary-use data files for the 1992 state and national reading data was issued in the spring of 1996.

Appendix H of the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* describes the error, its correction, and the revised results.

CHAPTER 3

INSTRUMENT DESIGN

3.1 Introduction

In the 1994 Trial State Assessment in reading, several types of instruments were used to collect data about students, teachers, and schools. Each assessed student received a booklet containing two segments of cognitive reading items, a demographic questionnaire, a reading background questionnaire, and a segment containing questions about the student's motivation. An excluded student questionnaire was completed by school officials for each sampled student who was deemed unable to take part in the assessment. Teacher questionnaires were given to the reading teachers of the assessed students. A school characteristics and policies questionnaire was distributed to each participating school to be completed by the school principal or other administrator.

This chapter describes the content and organization of the assessment instruments. See Chapter 4 for information about how schools, students, and teachers were selected to participate in the assessment.

3.2 Student Assessment Booklets

3.2.1 Booklet Content

The framework adopted for the 1994 reading assessment is organized according to a four-by-three matrix of reading *stances* by reading *purposes*. The stances included

- Initial Understanding,
- Developing an Interpretation,
- Personal Reflection and Response, and
- Demonstrating a Critical Stance.

These stances were assessed across three global purposes defined as

- Reading for Literary Experience,
- Reading to Gain Information, and
- Reading to Perform a Task.

Different types of texts were used to assess the various purposes for reading. Students' reading abilities were evaluated in terms of a single purpose for each type of text. At grade 4 only Reading for Literary Experience and Reading to Gain Information were assessed, while all three global purposes were assessed at grades 8 and 12. Figures 3-1 and 3-2 describe the four reading stances and three reading purposes that guided the development of the 1994 Trial State Assessment in reading. The distribution of items by reading purpose across grade levels is provided in Table 3-1. Table 3-2 shows the distribution of items by reading stance, as specified in the reading framework, for all three grade levels.

The development of cognitive items began with a careful selection of grade-appropriate passages for the assessment. Passages were selected from a pool of reading selections contributed by teachers from across the country. The framework stated that the assessment passages should represent authentic, naturally occurring reading material that students may encounter in and out of school. Furthermore, these passages were to be reproduced in test booklets as they had appeared in their original publications. Final passage selections were made by the Reading Instrument Development Panel. Finally, in order to guide the development of items, passages were outlined or mapped to identify essential elements of the text.

Figure 3-1
Description of Reading Stances

Readers interact with text in various ways as they use background knowledge and understanding of text to construct, extend, and examine meaning. The NAEP reading assessment framework specified four reading stances to be assessed that represent various interactions between readers and texts. These stances are not meant to describe a hierarchy of skills or abilities. Rather, they are intended to describe behaviors that readers at all developmental levels should exhibit.

Initial Understanding

Initial understanding requires a broad, preliminary construction of an understanding of the text. Questions testing this aspect ask the reader to provide an initial impression or unreflected understanding of what was read. In the 1992 and 1994 NAEP reading assessments, the first question following a passage was usually one testing initial understanding.

Developing an Interpretation

Developing an interpretation requires the reader to go beyond the initial impression to develop a more complete understanding of what was read. Questions testing this aspect require a more specific understanding of the text and involve linking information across parts of the text as well as focusing on specific information.

Personal Reflection and Response

Personal reflection and response requires the reader to connect knowledge from the text more extensively with his or her own personal background knowledge and experience. The focus is on how the text relates to personal experience; questions on this aspect ask the readers to reflect and respond from a personal perspective. For the 1992 and 1994 NAEP reading assessments, personal response questions were typically formatted as constructed-response items to allow for individual possibilities and varied responses.

Demonstrating a Critical Stance

Demonstrating a critical stance requires the reader to stand apart from the text, consider it, and judge it objectively. Questions on this aspect require the reader to perform a variety of tasks such as critical evaluation, comparing and contrasting, application to practical tasks, and understanding the impact of such text features as irony, humor, and organization. These questions focus on the reader as interpreter/critic and require reflection and judgments.

Figure 3-2
Description of Purposes for Reading

Reading involves an interaction between a specific type of text or written material and a reader, who typically has a purpose for reading that is related to the type of text and the context of the reading situation. The 1994 NAEP reading assessment presented three types of text to students representing each of three reading purposes: literary text for literary experience, informational text to gain information, and documents to perform a task. Students' reading skills were evaluated in terms of a single purpose for each type of text.

Reading for Literary Experience

Reading for literary experience involves reading literary text to explore the human condition, to relate narrative events with personal experiences, and to consider the interplay in the selection among emotions, events, and possibilities. Students in the NAEP reading assessment were provided with a wide variety of literary text, such as short stories, poems, fables, historical fiction, science fiction, and mysteries.

Reading to Gain Information

Reading to gain information involves reading informative passages in order to obtain some general or specific information. This often requires a more utilitarian approach to reading that requires the use of certain reading/thinking strategies different from those used for other purposes. In addition, reading to gain information often involves reading and interpreting adjunct aids such as charts, graphs, maps, and tables that provide supplemental or tangential data. Informational passages in the NAEP reading assessment included biographies, science articles, encyclopedia entries, primary and secondary historical accounts, and newspaper editorials.

Reading to Perform a Task

Reading to perform a task involves reading various types of materials for the purpose of applying the information or directions in completing a specific task. The reader's purpose for gaining meaning extends beyond understanding the text to include the accomplishment of a certain activity. Documents requiring students in the NAEP reading assessment to perform a task included directions for creating a time capsule, a bus schedule, a tax form, and instructions on how to write a letter to a senator. In 1994, reading to perform a task was assessed only at grades 8 and 12.

Table 3-1
Percentage Distribution of Items by Grade and Reading Purpose

Grade	Purposes for Reading: Reading for Literary Experience	Reading to Gain Information	Reading to Perform a Task
4	55%	45%	(No Scale)
8	40%	40%	20%
12	35%	45%	20%

Table 3-2
Percentage Distribution of Items by Reading Stance for Grades 4, 8, and 12

Initial Understanding/ Developing an Interpretation	Personal Reflection and Response	Demonstrating a Critical Stance
33%	33%	33%

The Trial State Assessment included constructed-response (short and extended) and multiple-choice items. The decision to use a specific item type was based on a consideration of the most appropriate format for assessing the particular objective. Both types of constructed-response items were designed to provide an in-depth view of students' ability to read thoughtfully and generate their own responses to reading. Short constructed-response questions, which were scored correct/incorrect, were used when students needed to respond in only one or two sentences in order to demonstrate full comprehension. Extended constructed-response questions, which were scored on a partial credit scale, were used when the task required more thoughtful consideration of the text and engagement in more complex reading processes. Multiple-choice items were used when a straightforward, single correct answer was all that was required. Guided by the NAEP reading framework, the Instrument Development Panel monitored the development of all three types of items to assess objectives in the framework.

The Trial State Assessment included eight different 25-minute "blocks," each consisting of one or more passages and a set of multiple-choice and constructed-response items to assess students' comprehension of the written material. Students were asked to respond to two 25-minute blocks within one booklet.

The overall pool of cognitive items for the Trial State Assessment in reading consisted of 84 items, including 37 short constructed-response items, 8 extended constructed-response items, and 39 multiple-choice items.

In addition to the cognitive items, students were asked a set of questions about their demographic characteristics, a set of questions about their reading background, and a set of questions about their motivation. These questionnaires are described in Section 3.3.1.

Tables 3-3 and 3-4 provide the composition of each block of items administered in the Trial State Assessment Program in reading.

3.2.2 Booklet Assembly

Each student assessment booklet included two sections of cognitive reading items and three sections of background questions. The assembly of reading blocks into booklets and their subsequent assignment to sampled students was determined by a *partially balanced incomplete block* (PBIB) design with *spiraled* administration.

The first step in implementing PBIB spiraling for the grade 4 reading assessment required constructing blocks of passages and items that required 25 minutes to complete. These blocks were then assembled into booklets containing two 5-minute background sections, one 3-minute background section, and two 25-minute blocks of reading passages and items according to a partially balanced incomplete block design. The overall assessment time for each student was approximately 63 minutes.

At the fourth-grade level, the blocks measured two purposes for reading—reading for literary experience and reading to gain information. The reading blocks were assigned to booklets in such a way that every block within a given purpose for reading was paired with every other block measuring the same purpose but was only paired with one block measuring the other purpose for reading. Every block appears in four booklets—three times within booklets measuring the same purpose and once in a booklet measuring both purposes. This is the *partially balanced* part of the balanced incomplete block design.

The PBIB design for both the 1992 and 1994 national reading assessment (and also for the Trial State Assessments) was *focused*—each block was paired with every other reading block assessing the same purpose for reading but not with all the blocks assessing the other purpose for reading. The *focused*-PBIB design also balances the order of presentation of the blocks of items—every block appears as the first cognitive block in two booklets and as the second cognitive block in two other booklets.

The design used in 1994 required that eight blocks of grade 4 reading items be assembled into 16 booklets. The assessment booklets were then *spiraled* and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the PBIB-spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, students in an assessment session received different booklets, and only a few students in a session received the same booklet. In the Trial State Assessment design, across all jurisdictions, representative and randomly equivalent samples of about 25,625 students responded to each item.

Table 3-3 provides the composition of each block of items administered in the Trial State Assessment Program in reading. Table 3-4 provides the total number of booklets, cognitive blocks, and noncognitive blocks used for the program. Table 3-4 also provides the details of the focused-PBIB design that was used with 8 blocks and 16 booklets.

3.2.3 Release Status for Item Blocks

As described in Section 1.4, "Item Security," some NAEP cognitive items are available for unrestricted public use. In the 1994 Trial State Assessment data files, all items in the cognitive block R3 are classified as public release and are available to secondary users.

All student demographic and reading background items and items from teacher, school, and excluded student questionnaires are also classified as public release and are available to secondary users.

Table 3-3
Cognitive and Noncognitive Block Information

Block	Type	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items	Booklets Containing Block
B1	Common Background	22	22	0	30 - 45
R2	Reading Background	15	15	0	30 - 45
RB	Reading Motivation	5	5	0	30 - 45
R3	Reading for Literary Experience	11	6	5	30, 31, 35, 43
R4	Reading for Literary Experience	12	5	7	30, 33, 34, 42
R5	Reading for Literary Experience	11	7	4	31, 32, 34, 44
R6	Reading to Gain Information	10	5	5	36, 39, 40, 44
R7	Reading to Gain Information	10	4	6	37, 38, 40, 42
R8*	Reading to Gain Information	9	3	6	38, 39, 41, 43
R9*	Reading for Literary Experience	9	3	6	32, 33, 35, 45
R10	Reading to Gain Information	12	6	6	36, 37, 41, 45

Note: *New blocks for the 1994 assessment.*

Table 3-4
Booklet Contents

Booklet Number	Common Background Block	Cognitive Blocks	Reading Background Block	Reading Motivation Block
R1	B1	R4, R3	R2	RB
R2	B1	R3, R5	R2	RB
R3	B1	R5, R9	R2	RB
R4	B1	R9, R4	R2	RB
R5	B1	R4, R5	R2	RB
R6	B1	R3, R9	R2	RB
R7	B1	R6, R10	R2	RB
R8	B1	R10, R7	R2	RB
R9	B1	R7, R8	R2	RB
R10	B1	R8, R6	R2	RB
R11	B1	R6, R7	R2	RB
R12	B1	R10, R8	R2	RB
R13	B1	R7, R4	R2	RB
R14	B1	R8, R3	R2	RB
R15	B1	R5, R6	R2	RB
R16	B1	R9, R10	R2	RB

3.3 Questionnaires

As part of the Trial State Assessment (as well as the national assessment), a series of questionnaires was used to collect information about assessed students, excluded students, reading teachers, and schools. The questionnaires are described in the following sections; sampling methods are described in Chapter 4.

3.3.1 Student Questionnaires

In addition to the cognitive questions, the 1994 Trial State Assessment included three student questionnaires. Two of these were five-minute sets of general and reading background questions designed to gather contextual information about students, their instructional and recreational experiences in reading, and their attitudes toward reading. The third, a three-minute questionnaire, was given to students at the end of each booklet to determine students' motivation in completing the assessment and their familiarity with assessment tasks. In order to ensure that all fourth-grade students understood the questions and had every opportunity to respond to them, the three questionnaires were read aloud by administrators as students read along and responded in their booklets.

The **student demographics (common core) questionnaire** (22 questions) included questions about race/ethnicity, language spoken in the home, mother's and father's level of education, reading materials in the home, homework, attendance, which parents live at home, and which parents work. This questionnaire was the first section in every booklet. In many cases the questions used were continued from prior assessments, so as to document changes in contextual factors that occur over time.

Three categories of information were represented in the second five-minute section of reading background questions called the **student reading questionnaire** (14 questions): time spent studying reading (both the amount of instruction received in reading and the time spent on reading homework), instructional practices (related to reading in the classroom, including group work, special projects, and writing in response to reading), and attitudes towards reading (students were asked questions such as whether they enjoyed reading and whether they were good in reading). This questionnaire was the fourth section in each booklet.

The **student motivation questionnaire** (5 questions) asked students to describe how hard they tried on the NAEP reading assessment, how difficult they found the assessment, how many questions they thought they got right, how important it was for them to do well, and how familiar they were with the assessment format.

Data from these questionnaires are contained on the student data files.

3.3.2 IEP/LEP Student Questionnaire

The **IEP/LEP Student Questionnaire** was completed by the teachers of those students selected to participate in the assessment sample who had an Individualized Education Plan (IEP) or were classified as Limited English Proficient (LEP). The questionnaire was completed for all IEP or LEP students, whether or not they actually participated in the assessment. This questionnaire asked about the nature of the student's disability and the special programs in which the student participated.

Schools were permitted to exclude certain students from the assessment. In order to be excluded, a student must have 1) had an Individualized Education Plan and not been mainstreamed at least 50 percent of the time or 2) been categorized as Limited English Proficient. In addition, school staff would have had to have determined that it was inappropriate to include the student in the assessment.

Data from the IEP/LEP questionnaire for IEP/LEP students who took part in the assessment are contained on the student data files. Questionnaire data for IEP/LEP students excluded from the assessment are contained on the excluded student data files.

3.3.3 Teacher Questionnaire

To supplement the information on instruction reported by students, the reading teachers of the fourth graders participating in the Trial State Assessment were asked to complete a questionnaire about their instructional practices, teaching backgrounds, and characteristics. The teacher questionnaire contained two parts. The first part pertained to the teachers' background and general training. The second part pertained to specific training in teaching reading and the procedures the teacher uses for *each class* containing an assessed student.

The **Teacher Questionnaire, Part I: Background and General Training** (25 questions) included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields of study, coursework in education, coursework in specific subject areas, amount of in-service training, extent of control over instructional issues, and availability of resources for their classroom.

The **Teacher Questionnaire, Part II: Training in Reading and Classroom Instructional Information** (46 questions) included questions on the teacher's exposure to various issues related to reading and teaching reading through pre- and in-service training, ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, methods of assessing student progress in reading, instructional emphasis given to the reading abilities covered in the assessment, and use of particular resources.

Data collected from the teacher questionnaires are appended to the appropriate student records in the student data files.

***Note**: The purpose of this sample is to estimate the numbers of students whose teachers have various attributes, not to estimate the attributes of the teacher population. Because of the nature of the sampling for the Trial State Assessment, the responses to the reading teacher questionnaire do not necessarily represent all fourth-grade reading teachers in a state. Rather, they represent the teachers of the particular students being assessed.*

3.3.4 School Questionnaire

A **School Characteristics and Policies Questionnaire** was given to the principal or other administrator of each school that participated in the Trial State Assessment program. This information provided an even broader picture of the instructional context for students' reading achievement. This questionnaire (64 questions) included questions about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, drop-out rates, size and composition of teaching staff, policies about grouping students, curriculum, testing practices and uses, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

Data collected from the school questionnaires can be found on the school data files.

CHAPTER 4

SAMPLE SELECTION AND WEIGHTS

4.1 Introduction

This chapter describes the methods used by Westat, Inc., the survey contractor, to select the samples for the states participating in the 1994 Trial State Assessment (Section 4.2) and provides an overview of the sampling weights on the data files and how they were derived for the state samples (Section 4.3). A discussion of how to use the sampling weights is given in Chapter 8. Sampling and weighting procedures for the national portion of the assessment are described in the technical report for the 1994 national assessment.

4.2 Sample Selection

The target population for the 1994 Trial State Assessment Program included students in public and nonpublic schools[2] who were enrolled in the fourth grade at the time of assessment. The sampling frame included public and nonpublic schools having the relevant grade in each jurisdiction. The samples were selected based on a two-stage sample design: selection of schools within participating jurisdictions, and selection of students within schools. The first-stage samples of schools were selected with probability proportional to the fourth-grade enrollment in the schools. Special procedures were used for jurisdictions with many small schools, and for jurisdictions having small numbers of grade-eligible schools.

The sampling frame for each jurisdiction was first stratified by urbanization status of the area in which the school was located. The urbanization classes were defined in terms of large or midsize central city, urban fringe of large or midsize city, large town, small town, and rural areas. Within urbanization strata, schools were further stratified explicitly on the basis of minority enrollment in those jurisdictions with substantial Black or Hispanic student population. Minority enrollment was defined as the total percent of Black and Hispanic students enrolled in a school. Within minority strata, schools were sorted by median household income of the ZIP code area where the school was located.

A systematic random sample of about 100 fourth-grade schools was drawn with probability proportional to the fourth-grade enrollment of the school from the stratified frame of schools within each jurisdiction. Each selected school provided a list of eligible enrolled students, from which a systematic sample of students was drawn.

One session of 30 students was sampled within each school, except in Delaware, where as many as three sessions were sampled within a given school. The number of sessions (i.e., multiples of 30 students) selected in each Delaware school was proportional to the fourth-grade enrollment of the school. Overlap between the 1994 state and national samples was minimized.

Guidelines for school and student participation rates in public- and nonpublic-school samples were established to preempt publication of results from jurisdictions for which participation rates suggested the possibility of appreciable nonresponse bias. Table 4-1 provides participation status separately for public- and nonpublic-school samples for each jurisdiction.

For jurisdictions that had participated in the 1992 Trial State Assessment, 25 percent of their selected public schools were designated at random to be monitored during the assessment so that reliable comparisons could be made between sessions administered with and without monitoring. For jurisdictions that had not participated in the previous assessment, 50 percent of their selected public schools were designated to be monitored. Fifty percent of all nonpublic schools were designated to be monitored, regardless of whether or not the jurisdiction had previously participated.

[2]Nonpublic schools include parochial schools, private schools, Bureau of Indian Affairs schools, and domestic Department of Defense Education Activity schools. Special education schools are not included.

Table 4-1
1994 Trial State Assessment Participation

Jurisdiction	Public School	Nonpublic School
Alabama	YES	YES
Arizona	YES	YES[5]
Arkansas	YES	YES
California	YES	YES[3]
Colorado	YES	YES[4]
Connecticut	YES	YES[4]
Delaware	YES	YES[4]
District of Columbia	*	*
DoDEA Overseas	YES	**
Florida	YES	YES[3]
Georgia	YES	YES[4]
Guam	YES	YES
Hawaii	YES	YES[4]
Idaho	YES[1]	YES
Indiana	YES	YES
Iowa	YES	YES
Kentucky	YES	YES[4]
Louisiana	YES	YES[4]
Maine	YES	YES
Maryland	YES	YES[3]
Massachusetts	YES	YES
Michigan	YES[1]	NO
Minnesota	YES	YES
Mississippi	YES	YES[3]
Missouri	YES	YES
Montana	YES[2]	YES[3]
Nebraska	YES[2]	YES[3]
New Hampshire	YES[2]	YES[5]
New Jersey	YES	YES[4]
New Mexico	YES	YES
New York	YES	YES[3]
North Carolina	YES	YES[5]
North Dakota	YES	YES
Pennsylvania	YES[2]	YES[4]
Rhode Island	YES[2]	YES
South Carolina	YES	YES[3]
Tennessee	YES[2]	YES[5]
Texas	YES	YES[5]
Utah	YES	YES[5]
Virginia	YES	YES[4]
Washington	YES	NO
West Virginia	YES	YES
Wisconsin	YES[2]	YES[3]
Wyoming	YES	NO

**Note*: *Participated but did not release results.*

***Note*: *DoDEA nonpublic school data do not exist because all non-domestic schools are considered public schools.*

[1] Did not meet the overall public school participation rate guidelines.

[2] Did not satisfy one of the guidelines for public school sample participation rates.

[3] Did not meet the overall nonpublic school participation rate guidelines.

[4] Did not satisfy one of the guidelines for nonpublic school sample participation rates.

[5] Weights were not calculated for nonpublic school data, owing to insufficient numbers of either cooperating nonpublic schools or assessed nonpublic school students in the given jurisdiction. See Appendix D for more information.

The following sections provide some details of the various aspects of selecting the sample for the 1994 Trial State Assessment, including frame construction, the stratification process, updating the school frame with new schools, and the actual sample selection. A fuller discussion of sample selection, including details of school and student participation and exclusion rates for each jurisdiction, is given in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

4.2.1 Selection of Schools

4.2.1.1 Frame Construction

In order to draw the school samples for the 1994 Trial State Assessment, it was necessary to obtain a comprehensive list of public and nonpublic schools in each jurisdiction. For each school, Westat needed useful information for stratification purposes, reliable information about grade span and enrollment, and accurate information for identifying the school to the state coordinator (district membership, name, address).

Based on experience with the 1992 Trial State Assessment, and national assessments from 1984 to 1992, Westat elected to use the file made available by Quality Education Data, Inc. (QED). They used the National Center for Education Statistics' Common Core of Data (CCD) school file to check the completeness of the QED file. The QED file was missing minority and urbanization data for a sizable minority of schools (due to the inability of QED to match these schools with the corresponding CCD file). Considerable efforts were undertaken to obtain these variables for all schools in states where these variables were to be used for stratification.

4.2.1.2 Stratification

Selection of schools within participating states involved two stages of explicit stratification and one stage of implicit stratification. The two explicit stages for public schools were urbanization and minority enrollment. The two explicit stages for nonpublic schools were metro status and school types. The final stage for both public and nonpublic schools was median income.

Urbanization Classification

The NCES "type of location" variable was used to stratify fourth-grade schools into seven different urbanization levels:

1) *Large Central City:* a central city of a Metropolitan Statistical Area (MSA) with a population greater than or equal to 400,000, or a population density greater than or equal to 6,000 persons per square mile;

2) *Midsize Central City:* a central city of an MSA but not designated as a large central city;.

3) *Urban Fringe of Large Central City:* a place within an MSA of a large central city and defined as urban by the U.S. Bureau of Census;

4) *Urban Fringe of Midsize Central City:* a place within an MSA of a midsize central city and defined as urban by the U.S. Bureau of Census;

5) *Large Town:* a place not within an MSA, but with a population greater than or equal to 25,000, but less than 50,000, and defined as urban by the U.S. Bureau of Census;

6) *Small Town:* a place not within an MSA, but with a population less than 25,000, but greater than 2,499, and defined as urban by the U.S. Bureau of Census;

7) *Rural:* a place with a population of less than 2,500 and defined as rural by the U.S. Bureau of the Census.

The urbanization strata were created by collapsing type of location categories. The nature of the collapsing varied across jurisdictions and grades. Each urbanization stratum included a minimum of 10 percent of eligible students in the participating jurisdiction.

Minority Classification

The second stage of stratification was minority enrollment. Minority enrollment strata were formed within urbanization strata, based on the percentages of Black and Hispanic students. The three cases that occur are described in the following paragraphs.

Case 1: Urbanization strata with less than 10 percent Black students and 7 percent Hispanic students were not stratified by minority enrollment.

Case 2: Urbanization strata with greater than or equal to 10 percent Black students or 7 percent Hispanic students, but not more than 20 percent of each, were stratified by ordering percent minority enrollment within the urbanization classes and dividing the schools into three groups with about equal numbers of students per minority group.

Case 3: In urbanization strata with greater than 20 percent of both Black and Hispanic students, minority strata were formed with the objective of providing equal strata with emphasis on the minority group (Black or Hispanic) with the higher concentration. The stratification was performed as follows. The minority group with the higher percentage gave the primary stratification variable; the remaining group gave the secondary stratification variable. Within urbanization class, the schools were sorted based on the primary stratification variable and divided into two groups of schools containing approximately equal numbers of students. Within each of these two groups, the schools were sorted by the secondary stratification variable and subdivided into two subgroups of schools containing approximately equal numbers of students. As a result, within urbanization strata there were four minority groups, low Black/low Hispanic, low Black/high Hispanic, high Black/low Hispanic, and high Black/high Hispanic.

The cutpoints in minority enrollment used to classify urbanization strata into these three cases were developed empirically. They ensure that there is good opportunity to stratify by race and ethnicity, without creating very small strata that would lead to sampling inefficiency.

The minority groups were formed solely for the purpose of creating efficient stratification design at this stage of sampling. These classifications were not directly used in analysis and reporting of the data, but acted to reduce sampling errors for achievement-level estimates.

Metro Status

All schools in the sampling frame were assigned metro status based on their FIPS county code and Census Bureau Metropolitan Statistical Area Definitions as of December 31, 1992. The field indicated if the school was located within a metropolitan area or not. This field was used as the first-stage stratification variable for nonpublic schools.

School Type

All nonpublic schools in the sampling frame were assigned a school type (Catholic or other nonpublic) based on their QED school type variable. This field was used as the second-stage stratification variable for nonpublic schools.

Median Household Income

Prior to the selection of the school samples, the schools were sorted by their primary and secondary stratification variables in a serpentine order. Within this sorted list, the schools were sorted, in serpentine order, by the median household income. This final stage of sorting resulted in implicit stratification of median income. The data on median household income were related to the ZIP code area in which the school is located. These data, derived from the 1990 Census, were obtained from Donnelly Marketing Information Services.

4.2.1.3 Selection of School Sample

Control of Overlap of School Samples for National Educational Studies

The issue of school sample overlap has been relevant in all rounds of NAEP in recent years. To avoid undue burden on individual schools, NAEP developed a policy for 1994 of avoiding overlap between national and state samples. This was to be achieved without unduly distorting the resulting samples by introducing bias or substantial variance. The procedure used was an extension of the method proposed by Keyfitz (1951) and is described in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

Selection of Schools in Small Jurisdictions

For jurisdictions with small numbers of public schools—specifically, Delaware, the District of Columbia, and Guam—all of the eligible fourth-grade public schools were included in the sample with certainty. This did not occur in any of the nonpublic school samples.

New School Selection

A district-level file was constructed from the fourth-grade school frame. The file was divided into a small districts file, consisting of those districts in which there were at most three schools on the aggregate frame and no more than one fourth-, one eighth-, and one twelfth-grade school. The remainder of districts were denoted as "large" districts.

A sample of large districts was drawn in each jurisdiction. All districts were selected in Delaware, the District of Columbia, Hawaii, and Rhode Island. The remaining jurisdictions in the file of large districts (eligible for sampling) were divided into two files within each jurisdiction. Two districts were selected per jurisdiction with equal probability among the smaller districts with combined enrollment of less than or equal to 20 percent of the jurisdiction's enrollment. From the rest of the file, eight districts were selected per jurisdiction with probability proportional to enrollment. The breakdown given above applied to all jurisdictions except Alaska and Nevada, where four and seven districts were selected with equal probability and six and three districts were selected with probability proportional to enrollment, respectively.

The 10 selected districts in each jurisdiction were then sent a listing of all their schools that appeared on the QED sampling frame, and were asked to provide information about new schools not included in the QED frame. These listings, provided by selected districts, were used as sampling frames for selection of new schools.

The determination as to how many new schools were selected and how the data from selected schools was weighted is discussed in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

Designating Schools to Be Monitored

One-fourth of the selected public schools were designated at random to be monitored during the assessment field period in all jurisdictions that had also participated in the 1992 Trial State Assessment. One-half of the selected public schools were designated to be monitored in jurisdictions that had not participated in the 1992 assessment—specifically Montana, Washington, and Department of Defense Education Activity Overseas. One-half of all nonpublic schools in every jurisdiction (regardless of 1992 participation) were designated to be monitored. The details of the implementation of the monitoring process in the field are given in Chapter 4 of the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

The purpose of monitoring a random quarter or half of the schools was to ensure that the procedures were being followed throughout each jurisdiction by the school and district personnel administering the assessments, and to provide data adequate for assessing whether there was a significant difference

in assessment results between monitored and unmonitored schools within each jurisdiction.

The following procedure was used to determine the sample of schools to be monitored. The initially selected schools were sorted in the order in which they were systematically selected. New schools from "large" districts were added to the sample at the end of the list in random order. The sorted schools were then paired, and one member of every other pair was assigned at random, with probability 0.5, to be monitored. One member of each pair was assigned to be monitored in jurisdictions requiring 50 percent monitoring of public schools as well as for all nonpublic school samples. If there was an odd number of schools, the last school was assigned monitor status as if it were part of a pair.

School Substitution

A substitute school was assigned to each sampled school (to the extent possible) prior to the field period through an automated substitute selection mechanism that used distance measures as the matching criterion. Two passes were made at the substitution; one assigning substitutes from outside the sampled school's district, and a second pass lifting this constraint. This strategy was instigated by the fact that most school nonresponse is really at the district level.

A distance measure was used in each pass and was calculated between each sampled school and each potential substitute. The distance measure was equal to the sum of four squared, standardized differences. The differences were calculated between the sampled and potential substitute school's estimated grade enrollment, median household income, percent Black enrollment and percent Hispanic enrollment. Each difference was squared and standardized to the population standard deviation of the component variable (e.g., estimated grade enrollment) across all fourth-grade schools and all jurisdictions. The potential substitutes were then assigned to sampled schools by order of increasing distance measure. An acceptance limit was put on the distance measure of 0.60. A given potential substitute was assigned to one and only one sampled school. Some sampled schools did not receive assigned substitutes (at least in the first pass) because the number of potential substitutes was less than the number of sampled schools or the distance measure for all remaining potential substitutes outside of the same district was greater than 0.60.

In the second pass, the different district constraint was lifted and the maximum distance allowed was raised to 0.75. This generally brought in a small number of additional assigned substitutes. Although the selected cut-off points of 0.60 and 0.75 on the distance measure were somewhat arbitrary, they were decided upon by reviewing a large number of listings beforehand and finding a consensus on the distance measures at which substitutes began to appear unacceptable.

4.2.2 Selection of Student Samples

Schools initially sent a complete list of students to a central location in November 1993. Schools were not asked to list students in any particular order, but were asked to implement checks to ensure that all fourth-grade students were listed. Based on the total number of students on this list, called the Student Listing Form, sample line numbers were generated for student sample selection. To generate these line numbers, the sampler entered the number of students on the form and the number of sessions into a calculator that had been programmed with the sampling algorithm. The calculator generated a random start that was used to systematically select the student line numbers (30 per session). Delaware was the only jurisdiction for which more than one session was conducted in a school. Up to three sessions were conducted in Delaware public schools, with the exact number of sessions being determined by the fourth-grade enrollment of each school. To compensate for new enrollees not on the Student Listing Form, extra line numbers were generated for a supplemental sample of new students.

After the student sample was selected, the administrator at each school identified students who were incapable of taking the assessment either because they had an Individualized Education Plan or because they were Limited English Proficient.

More details on the procedures for student exclusion are presented in the report on field procedures for the Trial State Assessment Program.

When the assessment was conducted in a school, a count was made of the number of nonexcluded students who did not attend the session. If this number exceeded three students, the school was instructed to conduct a make-up session to which were invited all students who were absent from the initial session.

4.3 Weighting Procedures

Following the collection of assessment and background data from and about assessed and excluded students, sampling weights and associated sets of replicate weights were derived. The sampling weights are needed to make valid inferences from the student samples to the respective populations from which they were drawn and should be used for all analyses, whether exploratory or confirmatory. Replicate weights are used in the estimation of sampling variance, through the procedure known as *jacknife repeated replication*. See Chapter 8 for information about how to use the sampling and replicate weights.

The following is an overview describing the weight variables on the data files and summarizing the methods use to derive them. Full details of the weighting procedures are given in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

4.3.1 Full-Sample Weights

Each student was assigned a weight to be used for making inferences about the state's students. This weight is known as the *full-sample* or *overall* weight.

The full-sample weight contained three components—a base weight, an adjustment for school nonparticipation, and an adjustment for student nonparticipation. These are described in a general way below; full details are given in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

The student base weight—the inverse of the overall probability of selection of the sampled student—incorporated the probability of selection of the student's school, and of the student within a school.

The student base weight was a product of the base weight of the school in which the student was enrolled (BASEWT), the school nonresponse adjustment factor (ADJFAC), and the within-school student weight (STUDWGT).

The student base weight was then adjusted for student-level nonparticipation. Weighting adjustments seek to reduce the potential for bias from such nonparticipation by increasing the weights of students from schools similar to those schools not participating (adjustment factor ADJFAC), and increasing the weights of students similar to those students from within participating schools who did not attend the assessment session (or a make-up) as scheduled (adjustment factor STUDNRF).

For excluded students, the procedures to calculate the base weight and school nonparticipation adjustment factor were the same as for assessed students from the same school. Excluded student nonresponse adjustments were calculated to account for the fact that an excluded student questionnaire was not returned for a small percentage of excluded students.

Either of two alternatively scaled weights can be used as the full-sample weight for analyses at the student level. The first of these, ORIGWT, has been scaled so that the sum of weights for all students in each jurisdiction estimates the total number of fourth-grade assessable students in that jurisdiction's public schools. The second of these, WEIGHT, is a proportional rescaling of ORIGWT, carried out so that the sum of WEIGHT across students and jurisdictions is equal to the total Trial State Assessment sample size across all jurisdictions (i.e., the total number of assessed students in the Trial State Assessment). Both weights should provide identical estimates of means, proportions,

correlations, and other statistics of interest in analyses within each state as well as analyses involving data from more than one jurisdiction.

The base weight assigned to a school was the reciprocal of the probability of selection of that school. The school base weight reflected the actual probability used to select the school from the frame, including the impact of avoiding schools selected for the NAEP national samples.

The final school weight, adjusted for nonparticipation, is named SCHWTF. This weight should be used in analyses of the school questionnaire data.

4.3.2 Replicate Weights

In addition to estimation weights, a set of replicate weights was provided for each student. These replicate weights are used in calculating the sampling errors of estimates obtained from the data, using the jackknife repeated replication method. Chapter 8 describes the method of using these replicate weights to estimate sampling errors. The methods of deriving these weights were aimed at reflecting the features of the sample design appropriately in each jurisdiction, so that when the jackknife variance estimation procedure is implemented as intended, approximately unbiased estimates of sampling variance result.

Replication estimates the variance of the full sample. This process involves repeatedly selecting portions of the sample to calculate the statistic of interest. The estimates that result are called replicate estimates. The variability among these calculated quantities is used to obtain the full sample variance.

The process of forming these replicate estimates involves first dividing the sample elements among a set of replicate groups, then using the pattern of replicate groups in a systematic fashion to apply replicate weights to the file.

Similar to the estimation weights, a set of replicate weights was derived for each student. The replicate weights (SRWT01-62) correspond to the full-sample weight ORIGWT. They are used for estimating the sampling errors of estimates derivedusing the full sample weights. These weights are designed to reflect the method of sampling schools, and account for the type of stratification used and whether or not the student's school was included in the sample with certainty. The method of sampling students within schools is also reflected, implicitly in the case of noncertainty schools and explicitly for schools included with certainty. These overall replicate weights also reflect the impact on sampling errors of the school- and student-level nonresponse adjustments applied to the full sample weights.

At the school level, the replicate weights SCHWT01-62 on the school data files should be used to estimate the variance for population estimates obtained using the school weight (SCHWTF).

4.3.3 Summary of Weights and Their Use

Table 4-2 gives a summary of the sample weights and replicate weights and the purposes for which they should be used. Chapter 8 provides a detailed discussion of how to use the weights in conducting analyses.

Table 4-2
Summary of Weights for the 1994 Trial State Assessment

Group	Sample Weight	Replicate Weights	Use
Assessed Students	ORIGWT	SRWT01-62	Student-level analyses comparing students within or across states
			Student-level analyses comparing students in state to students in nation
			Student-level analyses comparing students within or across states when comparing monitored and unmonitored sessions
Excluded Students	XWEIGHT	XRWT01-62	Excluded student analyses within or across states
Schools	SCHWTF	SCHWT01-62	School-level analyses within or across states
			School-level analyses between nation and states

CHAPTER 5

DATA COLLECTION, MATERIALS PROCESSING, PROFESSIONAL SCORING, AND DATABASE CREATION

5.1 Introduction

In addition to sample selection, Westat, Inc., was responsible for field administration and data collection for the 1994 Trial State Assessment. When data collection was completed, assessment instruments were sent to National Computer Systems for processing and scoring. The resulting data files were then sent to ETS, where they were transcribed to a database ready for analysis. This chapter provides an overview of these activities, which are described in detail in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

5.2 Data Collection and Field Administration

Data collection for the 1994 Trial State Assessment involved a collaborative effort between staff in the participating states and schools and the NAEP contractors, especially Westat, Inc., the field administration contractor. Between January 31 and February 25, 1994, Westat assessed the reading knowledge of over 121,000 students from approximately 4,700 public and nonpublic schools across the 44 participating jurisdictions. Westat's data collection responsibilities included selecting the sample of schools and students for each participating jurisdiction, developing administration procedures and manuals, training the state personnel who conducted the assessments, and conducting an extensive quality assurance program.

Each jurisdiction participating in the 1994 Trial State Assessment was asked to appoint a state coordinator who became the liaison between NAEP/Westat staff and the participating schools. At the school level, a local administrator was responsible for preparing for and conducting the assessment session in one or more schools. These individuals were usually school or district staff and were trained by Westat staff. In addition, Westat hired and trained six field managers and 46 state supervisors, one for each jurisdiction (two supervisors were hired for DoDEA overseas schools, one working in Europe and the other in the Far East).

Each field manager was responsible for working with the state coordinators of seven or eight jurisdictions and for overseeing assessment activities. The primary tasks assigned to field managers were to obtain information about cooperation and scheduing, ensure that arrangements for the assessments were set and assessment administrators identified, and schedule the assessment administrators training sessions.

The state supervisors were responsible for the training of the assessment administrators and the selection of the samples of students to be assessed. Westat also hired and trained an average of four quality control monitors in each jurisdiction to monitor the assessment sessions.

5.3 Materials Processing and Data Entry

Upon completion of each assessment session, field administration personnel shipped the assessment booklets and forms from the field to National Computer Systems in Iowa City for entry into computer files, professional scoring (see Section 5.4), checking, and creating the data files for transmittal to ETS. Over 175,000 booklets and questionnaires were received and processed for the Trial State Assessment in reading.

The student data and most of the questionnaire data were transcribed through the use of three separate systems:

- *data entry*, which included optical mark recognition scanning, image scanning, and intelligent character recognition;
- *validation* (edit); and
- *resolution*.

An intelligent data entry system was used for resolution of the scanned data, the entry of documents rejected by the scanning machine, and the entry of key-entered information. Additionally, each piece of input data was checked to verify that it was of an acceptable type, that it was within a specified range or ranges of values, and that it was consistent with other data values.

The introduction of image processing and image scoring further enhanced the work of NAEP. Image processing and scoring were successfully piloted in a side-by-side study conducted during the 1993 NAEP field test, and so became the primary processing and scoring methods for the 1994 Trial State Assessment. Image processing allowed the automatic collection of handwritten demographic data from the administrative schedules and the student test booklet covers through intelligent character recognition (ICR). This service was a benefit to the jurisdictions participating in NAEP because they were able to write rather than grid certain information—a significant reduction of burden on the schools. Image processing also made image scoring possible, eliminating much of the time spent moving paper. The images of student responses to be scored were transmitted electronically to the scoring center, located at a separate facility from where the materials were processed.

The high volume of collected data and the complexity of the Trial State Assessment processing design, with its spiraled distribution of booklets, as well as the concurrent administration of this assessment and the national assessments, required the development and implementation of flexible, innovative processing programs and a sophisticated process control system. This system allowed an integration of data entry and workflow management systems, including carefully planned and delineated editing, quality control, and auditing procedures.

5.4 Professional Scoring of Reading Items

The 1994 Trial State Assessment in reading contained three different types of cognitive items: extended constructed-response, short constructed-response, and multiple-choice. These items were administered in scannable assessment booklets that were identical to those used for the fourth-grade national assessment.

Scoring of the 1994 NAEP Trial State Assessment constructed-response items was conducted using NCS's image technology. All 1994 responses were scored online by readers working at image stations. The logistical problems associated with handling large quantities of student booklets were eliminated for those items scored on the image system.

One of the greatest advantages image technology presented for NAEP scoring was in the area of sorting and distributing work to scorers. All student responses for a particular item, regardless of where spiraling had placed that item in the various booklet forms, were grouped together for presentation to a team of 6 to 10 readers. This allowed training to be conducted one item at a time, rather than in blocks of related items, thus focusing readers' attention on the complexities of a single item.

Concerns about possible reader fatigue or other problems that might result from working continuously at a computer terminal proved unfounded. Both readers and table leaders responded with enthusiasm to the system, remarking on the ease with which student responses could be read and on the increased sense of professionalism they felt in working in this technological environment. Readers took periodic breaks, in addition to their lunch break, to reduce the degree of visual fatigue.

5.4.1 Description of Scoring

Each constructed-response item had a unique scoring standard that identified the range of possible scores for the item and defined the criteria to be used in evaluating the students' responses. The point values contained on the data files were assigned with the following meanings:

Dichotomous items from the 1992 assessment

- 1 = Unacceptable
- 2 = Acceptable

Dichotomous items developed during the 1993 field test

- 1 = Evidence of little or no comprehension
- 2 = Evidence of full comprehension

Three-point items developed during the 1993 field test

- 1 = Evidence of little or no comprehension
- 2 = Evidence of partial or surface comprehension
- 3 = Evidence of full comprehension

All four-point items

- 1 = Evidence of unsatisfactory comprehension
- 2 = Evidence of partial comprehension
- 3 = Evidence of essential comprehension
- 4 = Evidence of extensive comprehension

The scores for these items also included categories for no response, an erased, crossed-out, or illegible response, and any response found to be unrateable (i.e., off-task, responses written in a language other than English, or responses of "I don't know").

Figure 5-1 shows the scoring guide used for one of the extended constructed-response reading items.

A minimum of 25 percent of the 1994 reading responses were scored twice to determine reader reliability. The image system presented all responses in the same manner, so the reader could not discern which responses were being first-scored and which were designated for a second scoring. The table leader and the ETS trainer were able to monitor these figures on demand. The system showed the overall reliability for the group scoring the item and individual reliability of the qualified readers.

During the scoring of an item, the table leader could monitor progress using an interreader reliability tool. This display tool could be used in either of two modes—to display information of first readings versus second readings, or to display first reading of an individual versus second readings of that individual.

The table leaders were able to monitor work flow using a status tool that displayed the number of items completed, the number of items that still needed second scoring, and the number of items that had not been scored up to that time.

Table 5-1 shows the number of constructed-response items falling into each range of percentages of exact agreement.

Table 5-1
1994 NAEP Trial State Assessment
Number of Constructed-Response Items
in Each Range of Percentages of Exact Agreement Between Readers

Grade 4 Reading Items	Number of Unique Items	60-69%	70-79%	80-89%	90-100%
Short constructed-response items	37	0	0	8	29
Extended constructed-response items	8	0	1	6	1

Figure 5-1
Extended Constructed-Response Scoring Guide
for "Sybil Sounds the Alarm"

Question

What are the major events in the story [Sybil Sounds the Alarm]?

Stance

Initial Understanding

General Scoring Rubric

Demonstrates an understanding of an historical narrative by summarizing the important major events.

Unsatisfactory - These responses demonstrate little or no understanding of the events surrounding Sybil's ride by providing bits of information from the story, but not major events. In addition, these responses include those in which students merely copy one or more lines from the text, often the first or last sentence of the story.

Partial - These responses demonstrate some understanding of Sybil's ride by providing an account of one or two major events, not usually accompanied a detailed account or an explanation of the importance of the events. These responses may also be a brief statement without specific events.

Essential - These responses demonstrate an understanding of at least two of the major events surrounding Sybil's ride by providing a detailed account of these events **OR** by explaining the importance of the major events.

Extensive - These responses demonstrate an in-depth understanding of the major events surrounding Sybil's ride by providing a detailed account of major events accompanied by an explanation of their significance. The responses display a thorough understanding of the story as a whole.

No response (blank)

Crossed out, erased, or illegible

Not rateable (I don't know, Off task)

5.4.2 Constructed-Response Scores in the Secondary-Use Data Files

In the data file codebooks and layouts, constructed-response items that were dichotomized for scaling are identified by "OS" in the type field. Items that were scaled under a polytomous response model are identified by "OE" in the type field. The range of valid responses and the correct response(s) are given in the layouts in the RANGE and KEY VALUE fields.

5.5 Database Creation

The data transcription and editing procedures described above resulted in the generation of disk and tape files containing various data for assessed students, excluded students, teachers, and schools. The weighting procedures resulted in the generation of data files that included the sampling weights required to make valid statistical inferences about the population from which the 1994 fourth-grade Trial State reading assessment samples were drawn. These files were merged into a comprehensive, integrated database. To evaluate the effectiveness of the quality control of the data entry process, the corresponding portion of the final integrated database was verified in detail against the original instruments received from the field.

The transcription process conducted by NCS resulted in the transmittal to ETS of four data files: one file for each of the three questionnaires (teacher, school, and excluded student) and one for the student response data. The sampling weights, derived by Westat, Inc., comprised an additional three files—one for students, one for schools, and one for excluded students. (See Chapter 7 for a discussion of the sampling weights.) These seven files were the foundation for the analysis of the 1994 Trial State Assessment data. Before data analyses could be performed, these data files had to be integrated into a coherent and comprehensive database.

The 1994 Trial State Assessment database consisted of three files—student, school, and excluded student. Each record on the student file contained a student's responses to the particular assessment booklet the student was administered (Booklets R1 to R16) and the information from the questionnaire that the student's reading teacher completed. Additionally, for those assessed students who were identified as having an Individualized Education Plan (IEP) or Limited English Proficiency (LEP), data from the IEP/LEP Questionnaire is included. (Note that beginning with the 1994 assessment, the IEP/LEP questionnaire replaces the excluded student questionnaire. This questionnaire is filled out for all students identified as IEP and/or LEP, both assessed and excluded.) Since teacher response data can be reported only at the student level, it was not necessary to have separate teacher files. The school files and excluded student files were separate and could be linked to the student files through the state and school codes.

The creation of the student data files began with the reorganization of the data files received from NCS. This involved two major tasks: 1) the files were restructured, eliminating unused (blank) areas to reduce the size of the files; and 2) in cases where students had chosen not to respond to an item, the missing responses were recoded as either "omit" or "not reached," as appropriate. Next, the student response data were merged with the student weights file. The resulting file was then merged with the teacher response data. In both merging steps, the booklet ID (the two-digit booklet number and a five-digit serial number) was used as the matching criterion.

The school file was created by merging the school questionnaire file with the school weights file and a file of school variables, supplied by Westat, that included demographic information about the schools collected from the principal's questionnaire. The state and school codes were used as the matching criteria. Since some schools did not return a questionnaire and/or were missing principal's questionnaire data, some of the records in the school file contained only school-identifying information and sampling weight information.

The excluded student file was created by merging the excluded student questionnaire file with

the excluded student weights file. The assessment booklet serial number was used as the matching criterion.

When the student, school, and excluded student files had been created, the database was ready for analysis. In addition, whenever new data values, such as composite background variables or plausible values, were derived, they were added to the appropriate database files using the same matching procedures as described above.

To evaluate the effectiveness of the quality control of the data entry process, student data from the final integrated database was sampled, and the data were verified in detail against the original instruments received from the field. For this purpose, a number of student booklets were selected at random and compared, character by character, with their representation on the data files. The number of instruments involved in these quality control checks was based on the number needed to establish a statistically reassuring conclusion about the accuracy of the entire data entry operation. Results of the quality control checks are given in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading.*

CHAPTER 6

REPORTING SUBGROUPS AND OTHER VARIABLES

6.1 Introduction

In addition to overall achievement results, the 1994 Trial State Assessment permits reporting on the performance of various subpopulations of the student population. Some reporting subgroups were defined directly from responses to assessment items; some were derived from responses to two or more different items. Section 6.2 defines the reporting subgroups and explains how they are derived.

Certain variables on the data files were formed from the responses to one or more items from the student demographics questionnaire, the student reading background questionnaire, or the teacher questionnaire. These derived variables are described in Section 6.3.

Section 6.4 explains variables that were derived from students' responses to the reading items. Section 6.5 provides information about the proficiency variables (the plausible values) and other variables that were used in scaling student response data. Student and school file variables that come from the Quality Education Data, Inc. are explained in Section 6.6.

Values and response counts for all of the variables described in this chapter are found in the printed codebook for each state. Unless otherwise noted, the variables on the data files are named and defined in the same way for both the state sample and the national public-school sample that was used for state/nation comparisons.

6.2 Reporting Subgroups for the 1994 Trial State Assessment

Results for the 1994 Trial State Assessment were reported for student subgroups defined by gender, race/ethnicity, type of location, parents' level of education, and geographical region. The following explains how each of these subgroups was derived and the name of the variable to be used to perform secondary analyses of the subgroup data.

DSEX (Gender)

The variable SEX is the gender of the student being assessed, as taken from school records. For a few students, data for this variable was missing and was imputed by ETS after the assessment. The resulting variable DSEX contains a value for every student and is used for gender comparisons among students.

DRACE7 (Race/Ethnicity)

The variable DRACE7 is an imputed definition of race/ethnicity, derived from up to three sources of information. This variable is used for race/ethnicity subgroup comparisons. Two items from the student demographics questionnaire were used in the determination of derived race/ethnicity:

Demographic Item Number 2:

2. If you are Hispanic, what is your Hispanic background?

- ⬭ I am not Hispanic.
- ⬭ Mexican, Mexican American, or Chicano
- ⬭ Puerto Rican
- ⬭ Cuban
- ⬭ Other Spanish or Hispanic background

Students who responded to item number 2 by filling in the second, third, fourth, or fifth oval were considered Hispanic. For students who filled in the first oval, did not respond to the item, or provided information that was illegible or could not be classified, responses to item number 1 were examined in an effort to determine race/ethnicity. Item number 1 read as follows:

Demographic Item Number 1:

1. Which best describes you?

   - ○ White (not Hispanic)
   - ○ Black (not Hispanic)
   - ○ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish or Hispanic background.)
   - ○ Asian ("Asian" means someone who is Chinese, Japanese, Korean, Vietnamese, or other Asian background.)
   - ○ Pacific Islander ("Pacific Islander" means someone who is from a Filipino, Hawaiian, or other Pacific Island background.)
   - ○ American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)
   - ○ Other

Students' race/ethnicity was then assigned to correspond with their selection. For students who filled in the seventh oval ("Other"), provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity as provided from school records was used.

Derived race/ethnicity could not be determined for students who did not respond to background items 1 or 2 and for whom race/ethnicity was not provided by the school.

TOL8 (Type of Location)
TOL5
TOL3

The type of location variable was new for NAEP in 1994. It is closely based on a similar variable used by NCES in its Common Core of Data Public School Universe. This variable was used for three reasons. First, it seemed desirable to be consistent conceptually with other NCES data products. Second, the necessary data were available for each school to implement the code. Third, the classification is detailed (there are eight levels in all), thus giving maximum information and flexibility in reporting. The levels for type of location are:

1 Large Central City	a central city of a Metropolitan Statistical Area (MSA) or Primary Metropolitan Statistical Area (PMSA) with a population greater than or equal to 400,000, or a population density greater than or equal to 6,000 persons per square mile.
2 Midsize Central	a central city of a City MSA/PSA, but not a Large Central City
3 Urban Fringe of Large Central City	a place defined as urban that is within an MSA/PMSA that contains a Large Central City
4 Urban Fringe of Midsize City	a place defined as urban within a MSA/PMSA that contains no Large Central City
5 Large Town	a place not within an MSA/PMSA, with a population greater than or equal to 25,000
6 Small Town	a place not within a MSA, with a population less than 25,000, but greater than or equal to 2,500
7 Rural MSA	a place defined as rural (i.e., not within an Urbanized Area) within a MSA/PMSA
8 Rural NonMSA	a place not within a MSA/PMSA with a population of less than 2,500

These categories are designed to be exhaustive and mutually exclusive. Every place in the 50 United States and the District of Columbia is classified as belonging to exactly one of these categories. The definitions of MSAs and PMSAs, and their central cities, is carried out by the Office of Management and Budget (OMB). OMB Bulletin No. 93-17 states that "all agencies that conduct statistical activities to collect and publish data for Metropolitan Areas should use the most recent definitions of Metropolitan Areas established by OMB." The definitions used (as of June 30, 1993) were those current at the time of the 1994 assessment. The definitions of places and their populations are obtained from the published results of the 1990 Population Census, as are the definitions of Urbanized Areas.

Further details about the creation of the eight-category type of location variable are provided in *The NAEP 1994 Sampling and Weighting Report* (Wallace & Rust, 1996).

The variable TOL5 was created by collapsing the information provided in the variable TOL8 to five levels:

1 Large Central City
2 Midsize Central City
3 Urban Fringe of Large City, Urban Fringe of Midsize City, and Large Town
4 Small Town
5 Rural MSA and Rural NonMSA

The variable TOL3 is used extensively in the NAEP reports. TOL3 collapses TOL8 to three levels:

1	Central City	(Large Central City and Midsize Central City) This category includes central cities of all MSAs. Central City is a geographic term and is not synonymous with "inner city."
2	Urban Fringe of Large Central City	(Urban Fringe of Large City, Urban Fringe of Midsize City, and Large Town) An Urban Fringe includes all densely settled places and areas within MSAs 2
	Urban Fringe of Large Central City (continued)	that are classified as urban by the Bureau of the Census. A Large Town is defined as a place outside MSAs with a population greater than or equal to 25,000.
3	Rural/Small Town	(Small Town, Rural MSA, and Rural NonMSA) Rural includes all places and areas with a population of less than 2,500 that are classified as rural by the Bureau of the Census. A Small Town is defined as a place outside MSAs with a population of less than 25,000 but greater than or equal to 2,500.

PARED (Student's report of parents' education level)

The variable PARED is derived from responses to two questions, B003501 and B003601, in the student demographic questionnaire. Students were asked to indicate the extent of their mother's education (B003501—How far in high school did your mother go?) by choosing one of the following:

- She did not finish high school.
- She graduated from high school.
- She had some education after high school.
- She graduated from college.
- I don't know.

Students were asked to provide the same information about the extent of their father's education (B003601—How far in high school did your father go?) by choosing one of the following:

- He did not finish high school.
- He graduated from high school.
- He had some education after high school.
- He graduated from college.
- I don't know.

The information was combined into one parental education reporting category (PARED) as follows: If a student indicated the extent of

education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. For students who did not know the level of education for both parents or did not know the level of education for one parent and did not respond for the other, the parental education level was classified as unknown. If the student did not respond for both parents, the student was recorded as having provided no response.

REGION (Region of the Country)

States were grouped into four geographical regions—Northeast, Southeast, Central, and West—as shown in Table 6-1. All 50 states and the District of Columbia are listed. The part of Virginia that is included in the Washington, DC, metropolitan statistical area is included in the Northeast region; the remainder of the state is included in the Southeast region.

MODAGE (Modal Age)

The modal age (the age of most of the students in the grade sample) for the fourth-grade students is age 9. A value of 1 for MODAGE indicates that the student is younger than the modal age; a value of 2 indicates that the student is of the modal age; a value of 3 indicates that the student is older than the modal age.

6.3 Variables Derived from the Student and Teacher Questionnaires

Several variables were formed from the systematic combination of response values for one or more items from either the student demographic questionnaire, the student reading background questionnaire, or the teacher questionnaire.

HOMEEN2 (Home Environment—Articles [of 4] in the Home)

The variable HOMEEN2 was created from the responses to student demographic items B000901 (Does your family get a newspaper regularly?), B000903 (Is there an encyclopedia in your home?), B000904 (Are there more than 25 books in your home?), and B000905 (Does your family get any magazines regularly?). The values for this variable were derived as follows:

1	0-2 types	The student responded to at least two items and answered Yes to two or fewer.
2	3 types	The student answered Yes to three items.
3	4 types	The student answered Yes to four items.
4	Omitted	The student answered fewer than two items.

Table 6-1
NAEP Geographic Regions

NORTHEAST	SOUTHEAST	CENTRAL	WEST
Connecticut	Alabama	Illinois	Alaska
Delaware	Arkansas	Indiana	Arizona
District of Columbia	Florida	Iowa	California
Maine	Georgia	Kansas	Colorado
Maryland	Kentucky	Michigan	Hawaii
Massachusetts	Louisiana	Minnesota	Idaho
New Hampshire	Mississippi	Missouri	Montana
New Jersey	North Carolina	Nebraska	Nevada
New York	South Carolina	North Dakota	New Mexico
Pennsylvania	Tennessee	Ohio	Oklahoma
Rhode Island	Virginia	South Dakota	Oregon
Vermont	West Virginia	Wisconsin	Texas
Virginia			Utah
			Washington
			Wyoming

SINGLEP (How many parents live at home)

SINGLEP was created from items B005601 (Does either your mother or your stepmother live at home with you?) and B005701 (Does either your father or your stepfather live at home with you?). The values for SINGLEP were derived as follows:

1	2 parents at home	The student answered Yes to both items.
2	1 parent at home	The student answered Yes to B005601 and No to B005701, or Yes to B005701 and No to B005601.
3	Neither at home	The student answered No to both items.
4	Omitted	The student did not respond to or filled in more than one oval for one or both items.

TRUMAJ (Teacher undergraduate major - Reading)

Items T040701 and T040705 through T040710 in the teacher questionnaire (What were your undergraduate major fields of study?) were used to determine TRUMAJ as follows:

1	English/Reading	The teacher responded Yes to T040706 or T040707 (English, reading, and/or language arts).
2	Education	The teacher responded Yes to T040701 (education) and No to T040706 and T040707
3	Other	Any other response

TRGMAJ (Teacher graduate major - Reading)

Items T040801 and T040805 through T040811 in the teacher questionnaire (What were your

graduate major fields of study?) were used to determine TRGMAJ as follows:

1	English/Reading	The teacher responded Yes to T040807 or T040808 (English, reading and/or language arts)
2	Education	The teacher responded Yes to T040801 (education) and No to T040807 and T040808.
3	Other	The teacher responded Yes to T040805 (other), T040809 (geograph), and T040810 (history), or T040811 (social studies).
4	None	The teacher indicated (T040806) that he or she had no graduate-level study.

6.4 Variables Derived from Cognitive Items

BKSCOR (Booklet-level score)

The booklet-level score is a student-level score based on the sum of the number correct for dichotomous items plus the sum of the scores on the polytomous items, where the score for a polytomous item starts from 0 for the unacceptable category. Thus, for a 4-point extended constructed-response item, scores of "no response", "off-task", and "unsatisfactory" are assigned an item score of 0. Scores of "partial", "essential", and "extensive" are assigned item scores of 1, 2, and 3, respectively. The score is computed based on all cognitive items in an individual's assessment booklet.

LOGIT (Logit percent correct within booklet)

In order to compute the LOGIT score, a percent correct within booklet was first computed. This score was based on the ratio of the booklet score (BKSCOR) over the maximum booklet score. The percent correct score was set to .0001 if no items were answered correctly; if BKSCOR equaled the maximum booklet score, the percent correct score was set to .9999. A logit score, LOGIT, was calculate for each student by the following formula:

A logit score, LOGIT, was calculated within booklet for each student by the following formula:

$$\text{LOGIT} = \ln\left[\frac{PCTCOR}{1 - PCTCOR}\right]$$

LOGIT was then truncated to a value x, such that $-3 \le x \le 3$. After computing LOGIT for each student, the mean and standard deviation was calculated for each booklet as the first step in standardizing the logit scores. The standardized logit score, ZLOGIT, was then calculated for each student by the following formula:

$$\text{ZLOGIT} = \left[\frac{LOGIT - mean\ logit}{standard\ deviation}\right]$$

NORMIT (Normit Gaussian Score) SCHNORM (School-Level Mean Gaussian Score)

The normit score is a student-level Gaussian score based on the inverse normal transformation of the mid-percentile rank of a student's number-correct booklet score within that booklet. The normit scores were used to decide collapsing of variables, finalize conditioning coding, and check the results of scaling.

The number correct is based on the number of dichotomous items answered correctly plus the score obtained on extended constructed-response items.

$$\frac{CF(i)+CF(i-1)}{2N}$$

The mid-percentile rank is based on the formula:where CF(I) is the cumulative frequency at

I items correct and N is the total sample size. If I = 0 then

$$\frac{CF(0)+\frac{CF(1)}{2}}{2N}$$

A school-level normit, SCHNORM, was also created; this was the mean normit across all reading booklets administered in a school. These school-level mean normit scores were used in conditioning procedures to take into account differences in school proficiency. For each school, the weighted mean of the logits for the students in that school was calculated. Each student was then assigned that mean as his or her school-level mean logit score value.

6.5 Variables Related to Proficiency Scaling

Proficiency Score Variables

Item response theory (IRT) was used to estimate average proficiency for the nation and for various subpopulations, based on students' performance on the set of cognitive items they received. IRT provides a common scale on which performance can be reported for the nation and subpopulations, even when all students do not answer the same set of questions. This common scale makes it possible to report on relationships between students' characteristics (based on their responses to the background questions) and their overall performance in the assessment.

A scale ranging from 1 to 500 was created to report performance for each content area. A composite scale was created based on a weighted average of the purpose-for-reading scales, where the weight for each content area was proportional to the relative importance assigned to the content area as specified in the reading objectives.

Scale proficiency estimates were obtained for all students. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions to compute population statistics. Plausible values are not optimal estimates of individual proficiency; instead, they serve as intermediate values to be used in estimating population characteristics. Chapter 8 provides further details on the computation and use of plausible values.

The proficiency score (plausible value) variables are provided on the student data files for each of the scales and are named as shown in Table 6-2.

SMEANR, SMNR1 **(School mean score using first plausible value)**
SRANKR, SRNKR1 **(School rank using first plausible value)**
SRNK3R, SRK3R1 **(Top, middle, bottom third using first plausible value)**

A mean reading composite score (SMEANR on the student files, SMNR1 on the school files) was calculated using the first composite plausible value for each school. The mean composite score was based on the values from the scaling variable RRPCM1 and was calculated using the students' sampling weights. The schools were then ordered from highest to lowest mean score (SRANKR on the student files, SRNKR1 on the school files) within a sample using school-level weights—the school with the highest mean score was given a ranking of 1 and the equal to the number of schools in the sample.

These variables were then used in partitioning the schools within the national grade sample into three groups (top third, middle third, and bottom third) based on their ranking (SRNK3R on the student files, SRK3R1 on the school files).

SMEANRP, SMNR1P **(School mean score using first plausible value, public schools only)**
SRANKRP, SRNKR1P **(School rank using first plausible value, public schools only)**

SRNK3RP, SRK3R1P (Top, middle, bottom third, using first plausible value, public schools only)

These variables were computed in the same manner as SMEANR, SMNR1, SRANKR, SRNKR1, SRNK3R, and SRK3R1 for the subset of fourth-grade students who attended public schools.

SMNRn (School mean score using plausible values 2 through 5)

SRNKRn (School rank using plausible values 2 through 5)

SRK3Rn (Top, middle, bottom third using plausible values 2 through 5)

SMNRnP (School mean score using plausible values 2 through 5, public schools only)

SRNKRnP (School rank using plausible values 2 through 5, public schools only)

SRK3RnP (Top, middle, bottom third, using plausible values 2 through 5, public schools only)

School ranking results presented in the 1994 NAEP reports are based on the first plausible value. However, since there are four additional estimates of proficiency (plausible values) for each student, school ranking data were also created for those estimates. These school rank values were created using the same procedures described above, substituting proficiency variables RRPCM2 through RRPCM5 to compute the results. In the variable names, n denotes the plausible value 2, 3, 4, or 5.

Note that these variables are included only on the school file.

6.6 Quality Education Data Variables (QED)

The data files contain several variables obtained from information supplied by Quality Education Data, Inc. (QED). QED maintains and updates annually lists of schools showing grade span, total enrollment, instructional dollars per pupil, and other information for each school. These data variables are retained on both the school and student files and are identified in the data layouts by "(QED)" in the SHORT LABEL field.

Most of the QED variables are defined sufficiently in the data codebooks. Explanations of others are provided below.

ORSHPT and SORSHPT are the Orshansky Percentile, an indicator of relative wealth that specifies the percentage of school-age children in a district who fall below the poverty line.

IDP and SIDP represent, at the school district level, dollars per student spent for textbooks and supplemental materials.

ADULTED and SADLTED indicate whether or not adult education courses are offered at the school site.

URBAN and SURBAN define the school's urbanicity: urban (central city); suburban (area surrounding central city, but still located within the counties constituting the metropolitan statistical area); or rural (area outside any metropolitan statistical area).

Table 6-2
Scaling Variables for the 1994 Trial State Assessment Sample

Reading Scale	Data Variables
Reading for Literary Experience	RRPS11 to RRPS15
Reading to Gain Information	RRPS21 to RRPS25
Composite	RRPCM1 to RRPCM5

CHAPTER 7

NAEP SCALING PROCEDURES AND THEIR APPLICATION IN THE TRIAL STATE ASSESSMENT

7.1 Overview

The primary method by which results from the Trial State Assessment are disseminated is scale-score reporting. With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or a series of scales even when different students have been administered different items. This chapter presents an overview of the scaling methodologies employed in the analyses of the data from NAEP surveys in general and from the Trial State Assessment in reading in particular. Details of the scaling procedures specific to the Trial State Assessment are presented in Section 7.6.

7.2 Background

The basic information from an assessment consists of the responses of students to the items presented in the assessment. For NAEP, these items are constructed to measure performance on sets of objectives developed by nationally representative panels of learning area specialists, educators, and concerned citizens. Satisfying the objectives of the assessment and ensuring that the tasks selected to measure each goal cover a range of difficulty levels typically requires many items. The Trial State Assessment in reading required 84 items at grade 4. To reduce student burden, each assessed student was presented only a fraction of the full pool of items through multiple matrix sampling procedures.

The most direct manner of presenting the assessment results is to report separate statistics for each item. However, because of the vast amount of information, having separate results for each of the items in the assessment pool hinders the comparison of the general performance of subgroups of the population. Item-by-item reporting masks similarities in trends and subgroup comparisons that are common across items.

An obvious summary of performance across a collection of items is the average of the separate item scores. The advantage of averaging is that it tends to cancel out the effects of peculiarities in items that can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare more easily the general performances of subpopulations.

Despite their advantages, there are a number of significant problems with average item scores. First, the interpretation of these results depends on the selection of the items; the selection of easy or difficult items could make student performance appear to be overly high or low. Second, the average score is related to the particular items comprising the average, so that direct comparisons in performance between subpopulations require that those subpopulations have been administered the same set of items. Third, because this approach limits comparisons to average scores on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, direct estimates of parameters or quantities such as the proportion of students who would achieve a certain score across the items in the pool are not possible when every student is administered only a fraction of the item pool. While the mean average score across all items in the pool can be readily obtained (as the average of the individual item scores), statistics that provide distributional information, such as quantiles of the distribution of scores across the full set of items, cannot be readily obtained without additional assumptions.

These limitations can be overcome by the use of response scaling methods. If several items require similar skills, the regularities observed in response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. These variables include a respondent-specific variable, called proficiency, which quantifies a respondent's

tendency to answer items correctly (or, for multipoint items, to achieve a certain score) and item-specific variables that indicate characteristics of the item such as its difficulty, effectiveness in distinguishing between individuals with different levels of proficiency, and the chances of a very low proficiency respondent correctly answering a multiple-choice item. (These variables are discussed in more detail in the next section). When combined through appropriate mathematical formulas, these variables capture the dominant features of the data. Furthermore, all students can be placed on a common scale, even though none of the respondents takes all of the items within the pool. Using the common scale, it becomes possible to discuss distributions of proficiency in a population or subpopulation and to estimate the relationships between proficiency and background variables.

It is important to point out that any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. Every item in a NAEP survey is of interest and can provide useful information about what young Americans know and can do. The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose.

The scaling for the Trial State Assessment in reading was carried out separately within the two reading content areas specified in the framework for grade 4 reading. This scaling within subareas was done because it was anticipated that different patterns of performance might exist for these essential subdivisions of the subject area. The two content area scales correspond with two purposes for reading—Reading for Literary Experience and Reading to Gain Information. By creating a separate scale for each of these content areas, potential differences in subpopulation performance between the content areas are preserved.

The creation of a series of separate scales to describe reading performance does not preclude the reporting of a single index of overall reading performance—that is, an overall reading composite. A composite is computed as the weighted average of the two content area scales, where the weights correspond to the relative importance given to each content area as defined by the framework. The composite provides a global measure of performance within the subject area, while the constituent content area scales allow the measurement of important interactions within educationally relevant subdivisions of the subject area.

7.3 Scaling Methodology

This section reviews the scaling models employed in the analyses of data from the Trial State Assessment in reading and the 1994 national reading assessment, and the multiple imputation or "plausible values" methodology that allows such models to be used with NAEP's sparse item-sampling design. The reader is referred to Mislevy (1991) for an introduction to plausible values methods and a comparison with standard psychometric analyses, to Mislevy, Johnson and Muraki (1992) and Beaton and Johnson (1992) for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach. It should be noted that the imputation procedure used by NAEP is a mechanism for providing plausible values for proficiencies and not for filling in blank responses to background or cognitive variables.

The 84 reading items administered at grade 4 in the Trial State Assessment were also administered to fourth-grade students in the national reading assessment. However, because the administration procedures differed, the Trial State Assessment data were scaled independently from the national data. The national data also included results for students in grades 8 and 12. Details of the scaling of the Trial State Assessment and the subsequent linking to the results from the national reading assessment are provided in Section 7.6.

7.3.1 The Scaling Models

Three distinct scaling models, depending on item type and scoring procedure, were used in the analysis of the data from the Trial State Assessment. Each of the models is based on item response theory

(IRT; e.g., Lord, 1980). Each is a "latent variable" model, defined separately for each of the scales, which express respondents' tendencies to achieve certain scores (such as correct/incorrect) on the items contributing to a scale as a function of a parameter that is not directly observed, called proficiency on the scale.

A three-parameter logistic (3PL) model was used for the multiple-choice items (which were scored correct/incorrect). The fundamental equation of the 3PL model is the probability that a person whose proficiency on scale *k* is characterized by the *unobservable* variable θ_k will respond correctly to item *j*:

$$P(x_j = 1|\theta_k, a_j, b_j, c_j) = \quad (7.1)$$

$$c_j + \frac{(1 - c_j)}{1 + \exp[-1.7a_j(\theta_k - b_j)]}$$

$$\equiv P_{j1}(\theta_k) ,$$

where

- x_j is the response to item *j*, 1 if correct and 0 if not;
- a_j where $a_j > 0$, is the slope parameter of item *j*, characterizing its sensitivity to proficiency;
- b_j is the threshold parameter of item *j*, characterizing its difficulty; and
- c_j where $0 \le c_j < 1$, is the lower asymptote parameter of item *j*, reflecting the chances of students of very low proficiency selecting the correct option.

Further define the probability of an incorrect response to the item as

$$P_{j0} = P(x_j = 0|\theta_k, a_j, b_j, c_j) \quad (7.2)$$
$$1 - P_{j1}(\theta_k)$$

A two-parameter logistic (2PL) model was used for short constructed-response items, which were scored correct or incorrect. The form of the 2PL model is the same as equations (7.1) and (7.2) with the c_j parameter fixed at zero.

Thirty-nine multiple-choice and 45 constructed-response items were presented in the Trial State and grade 4 national assessments. Of the latter, 37 were short constructed-response items, nine of which were scored on a three-point scale and 28 of which were dichotomously scored. The remaining eight constructed-response items were scored on a five-point scale with potential scores ranging from 0 to 4. Items that are scored on a multipoint scale are referred to as polytomous items, in contrast with the multiple-choice and short constructed-response items, which are scored correct/incorrect and referred to as dichotomous items.

The polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model is the probability that a person with proficiency θ_k on scale *k* will have, for the *j*th item, a response x_j that is scored in the *i*th of m_j ordered score categories:

$$P(X_j = i|\theta_k, a_j, b_j, d_{j,1}, \ldots, d_{j,m_j-1}) =$$

$$\frac{\exp(\sum_{v=0}^{i} 1.7a_j(\theta_k - b_j + d_{j,v})}{\sum_{g=0}^{m_j-1} \exp(\sum_{v=0}^{g} 1.7a_j(\theta_k - b_j + d_{j,v}))} \quad (7.3)$$

$$\equiv P_{ji}(\theta_k)$$

where

- m_j is the number of categories in the response to item *j*
- x_j is the response to item *j*, with possibilities 0,1,...,m_j-1
- a_j is the slope parameter;

b_j is the item location parameter characterizing overall difficulty; and

$d_{j,i}$ is the category i threshold parameter (see below).

Indeterminacies in the parameters of the above model are resolved by setting $d_{j,0} = 0$ and setting

$$\sum_{i=1}^{m_j-1} d_{j,i} = 0.$$

Muraki (1992) points out that $b_j - d_{j,i}$ is the point on the θ_k scale at which the plots of $P_{ji-1}(\theta_k)$ and $P_{ji}(\theta_k)$ intersect and so characterizes the point on the θ_k scale above which the category i response to item j has the highest probability of incurring a change from response category i-1 to i.

When $m_j = 2$, so that there are two score categories (0,1), it can be shown that $P_{ji}(\theta_k)$ of equation (7.3) for i=0,1 corresponds respectively to $P_{j0}(\theta_k)$ and $P_{j1}(\theta_k)$ of the 2PL model equations (7.1) and 7.2 with c_j=0).

A typical assumption of item response theory is the conditional independence of the response by an individual to a set of items, given the individual's proficiency. That is, conditional on the individual's θ_k, the joint probability of a particular response pattern $\underline{x} = (x_1,...,x_n)$ across a set of n items is simply the product of terms based on (7.1), (7.2), and (7.3):

$$P(\underline{x}|\theta_k, \textit{item parameters}) = \prod_{j=1}^{n} \prod_{i=0}^{m_j-1} P_{ji}(\theta_k)^{u_{ji}} \qquad (7.4)$$

where $P_{ji}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_j is taken equal to 2 for the dichotomously scored items, and u_{ji} is an indicator variable defined by

$$u_{ji} = \begin{cases} 1 \text{ if response } x_j \text{ was in category } i \\ 0 \text{ otherwise.} \end{cases}$$

It is also typically assumed that response probabilities are conditionally independent of background variables $(\underline{y})$, given θ_k, or

$$P(\underline{x}|\theta_k, \textit{item parameters}, \underline{y}) = p(\underline{x}|\theta_k, \textit{item parameters}) \qquad (7.5)$$

After $\underline{x}$ has been observed, equation (7.4) can be viewed as a likelihood function, and provides a basis for inference about θ_k or about item parameters. Estimates of item parameters were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs, and which concurrently estimates parameters for all items (dichotomous and polytomous). The item parameters are then treated as known in subsequent calculations. The parameters of the items constituting each of the separate scales were estimated independently of the parameters of the other scales. Once items have been calibrated in this manner, a likelihood function for the scale proficiency θ_k is induced by a vector of responses to any subset of calibrated items, thus allowing θ_k-based inferences from matrix samples.

In all NAEP IRT analyses, missing responses at the end of each block of items a student was administered were considered "not-reached," and treated as if they had not been presented to the respondent. Missing responses to dichotomous items before the last observed response in a block were considered intentional omissions, and treated as fractionally correct at the value of the reciprocal of the number of response alternatives. These conventions are discussed by Mislevy and Wu (1988). With regard to the handling of not-reached items, Mislevy and Wu found that ignoring not-reached items introduces slight biases into item parameter estimation to the degree that not-reached items are present and speed is correlated with ability. With regard to omissions, they found that the method described above provides consistent limited-information likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than responding randomly.

Although the IRT models are employed in NAEP only to summarize performance, a number of checks are made to detect serious violations of the assumptions underlying the models (such as

conditional independence). When warranted, remedial efforts are made to mitigate the effects of such violations on inferences. These checks include comparisons of empirical and theoretical item response functions to identify items for which the IRT model may provide a poor fit to the data.

Scaling areas in NAEP are determined *a priori* by grouping items into content areas for which overall performance is deemed to be of interest, as defined by the frameworks developed by the National Assessment Governing Board. A proficiency scale θ_k is defined *a priori* by the collection of items representing that scale. What is important, therefore, is that the models capture salient information in the response data to effectively summarize the overall performance on the content area of the populations and subpopulations being assessed in the content area. NAEP routinely conducts differential item functioning (DIF) analyses to guard against potential biases in making subpopulation comparisons based on the proficiency distributions.

The local independence assumption embodied in equation (7.4) implies that item response probabilities depend only on θ and the specified item parameters, and not on the position of the item in the booklet, the content of items around an item of interest, or the test-administration and timing conditions. However, these effects are certainly present in any application. The practical question is whether inferences based on the IRT probabilities obtained via (7.4) are robust with respect to the ideal assumptions underlying the IRT model. Our experience with the 1986 NAEP reading anomaly (Beaton & Zwick, 1990) has shown that for measuring small changes over time, changes in item context and speededness conditions can lead to unacceptably large random error components. These can be avoided by presenting items used to measure change in identical test forms, with identical timings and administration conditions. Thus, we do *not* maintain that the item parameter estimates obtained in any particular booklet configuration are appropriate for other conceivable configurations. Rather, we assume that the parameter estimates are context-bound. (For this reason, we prefer common population equating to common item equating whenever equivalent random samples are available for linking.) This is the reason that the data from the Trial State Assessment were calibrated separately from the data from the national NAEP—since the administration procedures differed somewhat between the Trial State Assessment and the national NAEP, the values of the item parameters could be different.

7.3.2 An Overview of Plausible Values Methodology

Item response theory was developed in the context of measuring individual examinees' abilities. In that setting, each individual is administered enough items (often 60 or more) to permit precise estimation of his or her θ, as a maximum likelihood estimate $\hat{\theta}$, for example. Because the uncertainty associated with each θ is negligible, the distribution of θ, or the joint distribution of θ with other variables, can then be approximated using individuals' $\hat{\theta}$ values as if they were θ values.

This approach breaks down in the assessment setting when, in order to provide broader content coverage in limited testing time, each respondent is administered relatively few items in a scaling area. The problem is that the uncertainty associated with individual θs is too large to ignore, and the features of the $\hat{\theta}$ distribution can be seriously biased as estimates of the θ distribution. (The failure of this approach was verified in early analyses of the 1984 NAEP reading survey; see Wingersky, Kaplan, & Beaton, 1987.) Plausible values were developed as a way to estimate key population features consistently, and approximate others no worse than standard IRT procedures would. A detailed development of plausible values methodology is given in Mislevy (1991). Along with theoretical justifications, that paper presents comparisons with standard procedures, discussions of biases that arise in some secondary analyses, and numerical examples.

The following provides a brief overview of the plausible values approach, focusing on its implementation in the Trial State Assessment analyses.

Let $\underline{y}$ represent the responses of all sampled examinees to background and attitude questions, along with design variables such as school membership, and let $\underline{\theta}$ represent the vector of scale proficiency values. If $\underline{\theta}$ were known for all sampled examinees, it would be possible to compute a statistic $t(\underline{\theta},\underline{y})$—such as a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T. A function $U(\underline{\theta},\underline{y})$—e.g., a jackknife estimate—would be used to gauge sampling uncertainty, as the variance of t around T in repeated samples from the population.

Because the scaling models are latent variable models, however, $\underline{\theta}$ values are not observed even for sampled students. To overcome this problem, we follow Rubin (1987) by considering $\underline{\theta}$ as "missing data" and approximate $t(\underline{\theta},\underline{y})$ by its expectation given $(\underline{x},\underline{y})$, the data that actually were observed, as follows:

$$t^*(\underline{x},\underline{y}) = E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}]$$

$$= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}|\underline{x},\underline{y})\, d\underline{\theta} \ . \qquad (7.6)$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the item responses x_i, background variables y_i, and model parameters for sampled student i. These values are referred to as imputations in the sampling literature, and plausible values in NAEP. The value of $\underline{\theta}$ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from the respondent's conditional distribution. Rubin (1987) proposes that this process be carried out several times—multiple imputations—so that the uncertainty associated with imputation can be quantified. The average of the results of, for example, M estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of (8.6); the variance among them, B, reflects uncertainty due to not observing θ, and must be added to the estimated expectation of $U(\underline{\theta},\underline{y})$, which reflects uncertainty due to testing only a sample of students from the population. Section 7.5 explains how plausible values are used in subsequent analyses.

It cannot be emphasized too strongly that **plausible values are *not* test scores for *individuals*** in the usual sense. Plausible values are offered only as intermediary computations for calculating integrals of the form of equation (7.6), in order to estimate *population* characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar θ estimates of educational measurement that are in some sense optimal for each examinee (e.g., maximum likelihood estimates, which are consistent estimates of an examinee's θ, and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population): *Point estimates that are optimal for individual examinees have distributions that can produce decidedly nonoptimal (specifically, inconsistent) estimates of population characteristics* (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects.

7.3.3 Computing Plausible Values in IRT-Based Scales

Plausible values for each respondent i are drawn from the conditional distribution $p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma)$, where Γ and Σ are regression model parameters defined in this subsection. This subsection describes how, in IRT-based scales, these conditional distributions are characterized, and how the draws are taken. An application of Bayes' theorem with the IRT assumption of conditional independence produces

$$p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma) \propto P(x_i|\underline{\theta}_i,y_i,\Gamma,\Sigma)$$
$$p(\underline{\theta}_i|y_i,\Gamma,\Sigma) = P(x_i|\underline{\theta}_i)\, p(\underline{\theta}_i|y_i,\Gamma,\Sigma) \ , \qquad (7.7)$$

where, for vector-valued $\underline{\theta}_i$, $P(x_i|\underline{\theta}_i)$ is the product over scales of the *independent likelihoods* induced by responses to items within each scale, and $p(\underline{\theta}_i|y_i,\Gamma,\Sigma)$ is the multivariate—and generally nonindependent—*joint density* of proficiencies for the scales, conditional on the observed value y_i of

background responses, and the parameters Γ and Σ. The scales are determined by the item parameter estimates that constrain the population mean to zero and standard deviation to one. The item parameter estimates are fixed and regarded as population values in the computation described in this subsection.

In the analyses of the data from the Trial State Assessment and the data from the national reading assessment, a normal (Gaussian) form was assumed for $p(\underline{\theta}_i|y_i,\Gamma,\Sigma)$, with a common variance-covariance matrix, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the first 134 to 200 principal components of 482 selected main effects and two-way interactions of the complete vector of background variables. The included principal components will be referred to as the *conditioning variables*, and will be denoted y^c. (The complete set of original background variables used in the Trial State Assessment reading analyses are listed in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.) The following model was fit to the data within each state:

$$\underline{\theta} = \Gamma\, y^c + \underline{\varepsilon}\,, \qquad (7.8)$$

where $\underline{\varepsilon}$ is multivariately normally distributed with mean zero and variance-covariance matrix Σ. The number of principal components of the conditioning variables used for each state was sufficient to account for 90 percent of the total variance of the full set of conditioning variables (after standardizing each variable). As in regression analysis, Γ is a matrix each of whose columns is the *effects* for one scale and Σ is the matrix *variance-covariance of residuals* between scales. By fitting the model (7.8) separately within each state, interactions between each state and the conditioning variables are automatically included in the conditional joint density of scale proficiencies.

Maximum likelihood estimates of Γ and Σ, denoted by $\hat{\Gamma}$ and $\hat{\Sigma}$, are obtained from Sheehan's (1985) MGROUP computer program using the EM algorithm described in Mislevy (1985). The EM algorithm requires the computation of the mean, $\underline{\bar{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in (7.7). These moments are computed using higher order asymptotic corrections (Thomas, 1992).

After completion of the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled respondents. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|x_i,y_i)$ that fixes Σ at the value $\hat{\Sigma}$, (Thomas, 1992). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean, $\underline{\bar{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in equation (8.7) (i.e., $p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma)$) are computed using the same methods applied in the EM algorithm. In the third step, the $\underline{\theta}_i$ are drawn independently from a multivariate normal distribution with mean $\underline{\bar{\theta}}_i$ and variance Σ_i^p, approximating the distribution in (7.7). These three steps are repeated five times producing five imputations of $\underline{\bar{\theta}}_i$ for each sampled respondent.

7.4 NAGB Achievement Levels

Since its beginning, a goal of NAEP has been to inform the public about what students in American schools know and can do. While the NAEP scales provide information about the distributions of proficiency for the various subpopulations, they do not directly provide information about the meaning of various points on the scale. Traditionally, meaning has been attached to educational scales by norm-referencing—that is, by comparing students at a particular scale level to other students. Beginning in 1990, NAEP reports have also presented data using achievement levels. The reading achievement levels were developed and adopted by the National Assessment Governing Board (NAGB), as authorized by the NAEP legislation. The achievement levels describe selected points on the scale in terms of the types of skills that are or should be exhibited by students scoring at that level. The achievement level process

was applied to the 1992 national NAEP reading composite and the 1994 national scales were linked to the 1992 national scales. Since the Trial State Assessment scales were linked to the national scales in both years, the interpretations of the selected levels also apply to the Trial State Assessment in 1994.

NAGB has determined that achievement levels shall be the first and primary way of reporting NAEP results. Setting achievement levels is a method for setting standards on the NAEP assessment that identify what students *should* know and be able to do at various points on the reading composite. For each grade in the national assessment and, here, for grade 4 in the Trial State Assessment, four levels were defined—*basic*, *proficient*, *advanced*, and the region *below basic*. Based on initial policy definitions of these levels, panelists were asked to determine operational descriptions of the levels appropriate with the content and skills assessed in the reading assessment. With these descriptions in mind, the panelists were then asked to rate the assessment items in terms of the expected performance of marginally acceptable examinees at each of these levels. These ratings were then mapped onto the NAEP scale to obtain the achievement level cutpoints for reporting. Further details of the achievement level-setting process appear in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

7.5 Analyses

When survey variables are observed without error from every respondent, standard variance estimators quantify the uncertainty associated with sample statistics from the only source of uncertainty, namely the sampling of respondents. Item-level statistics for NAEP cognitive items meet this requirement, but scale-score proficiency values do not. The IRT models used in their construction posit an unobservable proficiency variable θ to summarize performance on the items in the subarea. The fact that θ values are not observed even for the respondents in the sample requires additional statistical analyses to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. As described above, Rubin's (1987) multiple imputations procedures were adapted to the context of latent variable models to produce the plausible values upon which many analyses of the data from the Trial State Assessment were based. This section describes how plausible values were employed in subsequent analyses to yield inferences about population and subpopulation distributions of proficiencies.

7.5.1 Computational Procedures

Even though one does not observe the θ value of respondent i, one does observe variables that are related to it: x_i, the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i, the respondent's answers to demographic and background variables. Suppose one wishes to draw inferences about a number $T(\underline{\theta},\underline{Y})$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that if θ values were observable, we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\underline{\theta},\underline{y})$ [where $(\underline{\theta},\underline{y}) \equiv (\theta_1,y_1,\ldots,\theta_N,y_N)$], and that we could estimate the variance in t around T due to sampling respondents by the function $U(\underline{\theta},\underline{y})$. Given that observations consist of (x_i,y_i) rather than $(\underline{\theta}_i,y_i)$, we can approximate t by its expected value conditional on $(\underline{x},\underline{y})$, or

$$t^*(\underline{x},\underline{y}) = E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}] = \int t(\underline{\theta},\underline{y})\, p(\underline{\theta} \mid \underline{x},\underline{y})\, d\underline{\theta}\,.$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\underline{\theta}_i|x_i,y_i)$, which are obtained for all respondents by the method described in Section 7.3.3. Let $\underline{\hat{\theta}}_m$ be the mth such vector of plausible values, consisting of a multidimensional value for the latent variable of each respondent. This vector is a plausible representation of what the true $\underline{\theta}$ vector might have been, had we been able to observe it.

The following steps describe how an estimate of a scalar statistic $t(\underline{\theta},\underline{y})$ and its sampling variance can be obtained from M (>1) such sets of plausible

values. (Five sets of plausible values are used in NAEP analyses of the Trial State Assessment.)

1) Using each set of plausible values $\hat{\underline{\theta}}_m$ in turn, evaluate t as if the plausible values were true values of $\underline{\theta}$. Denote the results $\hat{t}_m$, for $m=1,...,M$.

2) Using the jackknife variance estimator defined in Chapter 8, compute the estimated sampling variance of $\hat{t}_m$, denoting the result U_m.

3) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \frac{\hat{t}_m}{M}.$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} \frac{U_m}{M}.$$

5) Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^*)^2}{(M-1)}$$

6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\, B.$$

***Note**: Due to the excessive computation that would be required, NAEP analyses did not compute and average jackknife variances over all five sets of plausible values, but only on the first set. Thus, in NAEP reports, U^* is approximated by U_1.*

7.5.2 Statistical Tests

Suppose that if θ values were observed for sampled students, the statistic $(t - T)/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^* - T)/V^{1/2}$ is approximately t-distributed, with degrees of freedom given by

$$\nu = \frac{1}{\frac{f^2}{M-1} + \frac{(1-f)^2}{d}}$$

where f is the proportion of total variance due to not observing θ values:

$$f_M = (1+M^{-1})\, B_M / V_M.$$

When B is small relative to U^*, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This is the case with main NAEP reporting variables. If, in addition, d is large, the normal approximation can be used to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_m and U^* is a covariance matrix, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T-t^*)\, V^{-1}\, (T-t^*)'$ is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f:

$$f = (1+M^{-1})\, Trace\,(BV^{-1})/k\,.$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

7.5.3 Biases in Secondary Analyses

Statistics t^* that involve proficiencies in a scaled content area and variables included in the conditioning variables y^c are consistent estimates of the corresponding population values T. Statistics involving background variables y that were *not* conditioned on, or relationships among proficiencies from *different* content areas, are subject to asymptotic biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variables to the

variables that were conditioned on and to the proficiency of interest. That is, the large sample expectations of certain sample statistics need not equal the true population parameters.

The *direction* of the bias is typically to underestimate the effect of nonconditioned variables. For details and derivations see Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987, Section 10.3.5). For a given statistic t^* involving one content area and one or more nonconditioned background variables, the *magnitude* of the bias is related to the extent to which observed responses x account for the latent variable θ, and the degree to which the nonconditioned background variables are explained by conditioning background variables. The first factor—conceptually related to test reliability—acts consistently in that greater measurement precision reduces biases in *all* secondary analyses. The second factor acts to reduce biases in certain analyses but increase it in others. In particular,

- high shared variance between conditioned and nonconditioned background variables *mitigates* biases in analyses that involve only proficiency and nonconditioned variables, such as marginal means or regressions; and

- high shared variance *exacerbates* biases in regression coefficients of conditional effects for nonconditioned variables, when nonconditioned and conditioned background variables are analyzed jointly as in multiple regression.

The large number of background variables that have been included in the conditioning vector for the Trial State Assessment allows a large number of secondary analyses to be carried out with little or no bias, and mitigates biases in analyses of the marginal distributions of θ in nonconditioned variables. Kaplan and Nelson's analysis of the 1988 NAEP reading data (some results of which are summarized in Mislevy, 1991), which had a similar design and fewer conditioning variables, indicates that the potential bias for nonconditioned variables in multiple regression analyses is below 10 percent, and biases in simple regression of such variables is below 5 percent. Additional research (summarized in Mislevy, 1990) indicates that most of the bias reduction obtainable from conditioning on a large number of variables can be captured by instead conditioning on the first several principal components of the matrix of all original conditioning variables. This procedure was adopted for the Trial State Assessment by replacing the conditioning effects by the first K principal components, where K was selected so that 90 percent of the total variance of the full set of conditioning variables (after standardization) was captured. Mislevy (1990) shows that this puts an upper bound of 10 percent on the average bias for all analyses involving the original conditioning variables.

7.6 Scaling the 1994 Trial State Assessment Reading Data

This section describes some of the details of the analyses carried out in developing the Trial State Assessment reading scales. The procedures used were similar to those employed in the analysis of the 1992 Trial State Assessments in reading (Allen, Mazzeo, Isham, Fong, & Bowker, 1994) and the 1990 and 1992 Trial State Assessments in mathematics (Mazzeo, 1991 and Mazzeo, Chang, Kulick, Fong, & Grima, 1993) and are based on the philosophical and theoretical underpinnings of the NAEP scaling procedures described in previous sections of this chapter.

The first step in the analysis of the Trial State Assessment data involved conventional item and test analyses—for example, examinations of average proportions correct, average biserial correlations, and differential item functioning. These analyses are discussed in detail in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*. This section focuses on the four major steps in the *scaling* of the Trial State Assessment data:

- item response theory (IRT) scaling;

- estimation of state and subgroup proficiency distributions based on the "plausible values" methodology;

- linking of the 1994 Trial State Assessment scales to the corresponding scales from the 1994 national assessment; and

- creation of the Trial State Assessment reading composite scale.

An overview of each of these steps is provided in the following sections. The rationale for and details of the steps are given in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.

7.6.1 Item Response Theory (IRT) Scaling

Separate IRT-based scales were developed using the scaling models described in Section 7.3. Two scales were produced by separately calibrating the sets of items classified in each of the two content areas.

A single set of item parameters for each item was estimated and used for all jurisdictions. Item parameter estimation was carried out using a 25 percent systematic random sample of the students participating in the 1994 Trial State Assessment and included equal numbers of students from each participating jurisdiction, half from monitored sessions and half from unmonitored sessions. All students in the scaling sample were public-school students. The sample consisted of 28,072 students, with 638 students being sampled from each of the 44 participating jurisdictions.

7.6.2 Item Parameter Estimation

For each content area scale, item parameter estimates were obtained using the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs[3]. The program uses marginal maximum likelihood estimation procedures to estimate the parameters of the one-, two-, and three-parameter logistic models, and the generalized partial credit model described by Muraki (1992).

Multiple-choice items were dichotomously scored and were scaled using the three-parameter logistic model. Omitted responses to multiple-choice items were treated as fractionally correct, with the fraction being set to 1 over the number of response options. Short constructed-response items that were also in the 1992 assessment were dichotomously scored and scaled using the two-parameter logistic model. New short (regular) constructed-response items were scored on a three-point generalized partial credit scale. These items appear in blocks 8 and 9. Omitted responses to short constructed-response items were treated as incorrect.

There were a total of eight extended constructed-response items. Each of these items was also scaled using the generalized partial credit model. Four scoring levels were defined:

0 Unsatisfactory response or omitted;
1 Partial response;
2 Essential response; and
3 Extensive response.

Note that omitted responses were treated as the lowest possible score level. As stated earlier, not-reached and off-task responses were treated as if the item was not administered to the student. Table 7-1 provides a listing of the blocks, positions within the block, content area classifications, and NAEP identification numbers for all extended constructed-response items included in the 1994 assessment.

Bayes modal estimates of all item parameters were obtained from the BILOG/PARSCALE program. Prior distributions were imposed on item parameters with the following starting values: thresholds, normal [0,2]; slopes, log-normal [0,.5]; and asymptotes, two-parameter beta with parameter

[3]Late in the analysis process, an error was discovered in the PARSCALE program documentation. This error affected the reading results, including those reported in the April 1995 version of the *First Look* report. The analyses and report were subsequently redone. Appendix H of the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* describes the error, its correction, and the revised results.

values determined as functions of the number of response options for an item and a weight factor of 50. The locations (but not the dispersions) were updated at each program estimation cycle in accordance with provisional estimates of the item parameters.

As was done for the 1990 and 1992 Trial State Assessments in mathematics and for the 1992 Trial State Assessment in reading, item parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed (normal [0,1]) and a stable solution was obtained. Starting values for the item parameters were provided by item analysis routines. The parameter estimates from this initial solution were then used as starting values for a subsequent set of runs in which the subject ability distribution was freed and estimated concurrently with item parameter estimates. After each estimation cycle, the subject ability distribution was re-standardized to have a mean of zero and standard deviation of one. Item parameter estimates for that cycle were correspondingly linearly transformed.

During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items in the grade 4 item pools. These evaluations were conducted to identify misfitting items, which would be excluded from the final item pool making up the scales. Evaluations of model fit were based primarily on a graphical analysis. For binary-scored items, model fit was evaluated by examining plots of nonmodel-based estimates of the expected conditional (on θ) proportion correct versus the proportion correct predicted by the estimated item characteristic curve (see Mislevy & Sheehan, 1987, p. 302). For the extended constructed-response items, similar plots were produced for each item category characteristic curve.

In making decisions about excluding items from the final scales, a balance was sought between being too stringent, hence deleting too many items and possibly damaging the content representativeness of the pool of scaled items, and too lenient, hence including items with model fit poor enough to invalidate the types of model-based inferences made from NAEP results. Items that clearly did not fit the model were not included in the final scales; however, a certain degree of misfit was tolerated for a number of items included in the final scales.

Only one item in the assessment received special treatment in the scaling process in both the 1992 and 1994 assessments. The generalized partial credit model did not fit the responses to the extended constructed-response item R012111 well. For this Reading for Literary Experience item, which appeared in the eleventh position in block R4, the categories 0 and 1 were combined and the other categories were relabeled. Therefore the codings for the three scoring levels were defined:

0 Unsatisfactory, partial response, or omitted;
1 Essential response; and
2 Extensive response.

The IRT parameters for the items included in the Trial State Assessment are listed in Appendix B.

7.6.3 Estimation of State and Subgroup Proficiency Distributions

The proficiency distributions in each jurisdiction (and for important subgroups within each jurisdiction) were estimated by using the multivariate plausible values methodology and the corresponding MGROUP computer program (described in Section 7.3; see also Mislevy, 1991). The MGROUP program (Sheehan, 1985; Rogers, 1991), which was originally based on the procedures described by Mislevy and Sheehan (1987), was used in the 1990 Trial State Assessment of mathematics. The 1992 and 1994 Trial State Assessments used an enhanced version of MGROUP, based on modifications described by Thomas (1992), to estimate the fourth-grade proficiency distribution for each jurisdiction. As described in the previous chapter, MGROUP estimates proficiency distributions using information from students' item responses, students' background variables, and the item parameter estimates obtained from the BILOG/PARSCALE program.

Table 7-1
Extended Constructed-Response Items, 1994 Trial State Assessment in Reading

Block	Position In Block	Scale	NAEP ID
R3	6	Reading for Literary Experience	R012006
R4	11	Reading for Literary Experience	R012111
R5	7	Reading for Literary Experience	R012607
R6	4	Reading to Gain Information	R012204
R7	8	Reading to Gain Information	R012708
R8	4	Reading to Gain Information	R015804
R9	7	Reading for Literary Experience	R015707
R10	12	Reading to Gain Information	R012512

Separate conditioning models were estimated for each jurisdiction. If a jurisdiction had a nonpublic-school sample, students from that sample were included in this part of the analysis, and a conditioning variable differentiating between public- and nonpublic-school students was included. This resulted in the estimation of 44 distinct conditioning models. The background variables included in each jurisdiction's model (denoted *y* in Section 7.3) were principal component scores derived from the within-jurisdiction correlation matrix of selected main-effects and two-way interactions associated with a wide range of student, teacher, school, and community variables. There were no interaction terms between independent variables in the 1992 Trial State Assessment in reading. However, in the 1994 assessment, interaction terms between certain independent variables that might be included in reports were added to the conditioning model. As was done for the 1992 Trial State Assessment, a set of five multivariate plausible values was drawn for each student who participated in the 1994 Trial State Assessment in reading.[4]

As was the case in previous assessments, plans for reporting each jurisdiction's results required analyses examining the relationships between proficiencies and a large number of background variables. The background variables included student demographic characteristics (e.g., the race/ethnicity of the student, highest level of education attained by parents), students' perceptions about reading, student behavior both in and out of school (e.g., amount of television watched daily, amount of homework done each day), and a variety of other aspects of the students' background and preparation, the background and preparation of their teachers, and the educational, social, and financial environment of the schools they attended. If a jurisdiction had a nonpublic-school sample, type of school was included as a background variable.

To avoid biases in reporting results and to minimize biases in secondary analyses, it is desirable to incorporate measures of a large number of independent variables in the conditioning model.

[4]There was one exception to this—in the 1994 public-school sample from Georgia. One student had an anomalous pattern of background characteristics that did not fit the conditioning model. After close scrutiny of the data for this student, it was determined that this outlying observation should be deleted from the principal component and conditioning portions of the analysis and from the results.

When expressed in terms of contrast-coded main effects and interactions, the number of variables to be included totaled 482. These contrasts were the common starting point in the development of the conditioning models for each of the participating jurisdictions. (A listing of the complete set of variables included in the conditioning model is provided in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*.)

Because of the large number of these contrasts and the fact that, within each jurisdiction, some contrasts had zero variance, some involved relatively small numbers of individuals, and some were highly correlated with other contrasts or sets of contrasts, an effort was made to reduce the dimensionality of the predictor variables in each jurisdiction's MGROUP models. As was done for the 1990 and 1992 Trial State Assessments in mathematics and the 1992 Trial State Assessment in reading, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting separate sets of principal components (one set for each of the 44 jurisdictions) from the within-jurisdiction correlation matrices of the original contrast variables. The principal components, rather than the original variables, were used as the independent variables in the conditioning model. As was done for the previous assessments, the number of principal components included for each jurisdiction was the number required to account for approximately 90 percent of the variance in the original contrast variables. Research based on data from the 1990 Trial State Assessment in mathematics suggests that results obtained using such a subset of the components will differ only slightly from those obtained using the full set (Mazzeo, Johnson, Bowker, & Fong, 1992).

It is important to note that the proportion of variance accounted for by the conditioning model differs across scales within a jurisdiction, and across jurisdictions within a scale. Such variability is not unexpected for at least two reasons. First, there is no reason to expect the strength of the relationship between proficiency and demographics to be identical across all jurisdictions. In fact, one of the reasons for fitting separate conditioning models is that the strength and nature of this relationship may differ across jurisdictions. Second, the homogeneity of the demographic profile also differs across jurisdictions. As with any correlational analysis, the restriction of the range in the predictor variables will attenuate the relationship.

NAEP scales are viewed as summaries of consistencies and regularities that are present in item-level data. Such summaries should agree with other reasonable summaries of the item-level data. In order to evaluate the reasonableness of the scaling and estimation results, a variety of analyses were conducted to compare state-level and subgroup-level performance in terms of the content area scaled scores and in terms of the average proportion correct for the set of items in a content area. High agreement was found in all of these analyses which showed that there is an extremely strong relationship between the estimates of state-level performance in the scale-score and item-score metrics for both content areas.

7.6.4 Linking State and National Scales

A major purpose of the Trial State Assessment Program was to allow each participating jurisdiction to compare its 1994 results with the nation as a whole and with the region of the country in which that jurisdiction is located. For meaningful comparisons to be made between each of the Trial State Assessment jurisdictions and the relevant national sample, results from these two assessments had to be expressed in terms of a similar system of scale units.

The purpose of this section is to describe the procedures used to align the 1994 Trial State scales with their 1994 national counterparts. The procedures that were used are similar to the common population equating procedures employed to link the 1990 national and state mathematics scales (Mazzeo, 1991; Yamamoto & Mazzeo, 1992) and the 1992 national and state mathematics and reading scales (Allen, Mazzeo, Isham, Fong & Bowker, 1994; Mazzeo, Chang, Kulick, Fong, & Grima, 1993).

Using the sampling weights provided by Westat, the combined sample of students from participating jurisdictions (a total sample size of 112,153) was used to estimate the distribution of proficiencies for the population of students enrolled in public schools in the participating states and the District of Columbia[5]. Data were also used from a subsample of 5,063 students in the national assessment at grade 4, consisting of grade-eligible public-school students from jurisdictions that contributed students to the combined sample from the Trial State Assessment. Appropriate weights were provided by Westat to obtain estimates of the distribution of proficiency for the same target population.

Thus, for each of the two scales, two sets of proficiency distributions were obtained. One set, based on the sample of combined data from the Trial State Assessment (referred to as the Trial State Assessment Aggregate Sample) and using item parameter estimates and conditioning results from that assessment, was in the metric of the 1994 Trial State Assessment. The other, based on the sample from the 1994 national assessment (referred to as the State Aggregate Comparison, or SAC, sample) and obtained using item parameters and conditioning results from that assessment, was in the metric of the 1994 national assessment. The latter metric had already been set using procedures described in the technical report of the 1994 national assessment. The two Trial State Assessment and national scales were made comparable by constraining the mean and standard deviation of the two sets of estimates to be equal.

[5]Students from Guam and DoDEA overseas schools were excluded from the definition of this target population; hence, data from students from these jurisdictions were not included in the combined Trial State Assessment samples.

More specifically, the following steps were followed to linearly link the scales of the two assessments:

1) For each scale, estimates of the proficiency distribution for the Trial State Assessment Aggregate Sample were obtained using the full set of plausible values generated by the CGROUP program. The weights used were the final sampling weights provided by Westat, not the rescaled versions. For each scale, the arithmetic mean of the five sets of plausible values was taken as the overall estimated mean and the standard deviations of the five sets of plausible values was taken as the overall estimated standard deviation.

2) For each scale, the estimated proficiency distribution of the State Aggregate Comparison sample was obtained, again using the full set of plausible values generated by the CGROUP program. The weights used were specially provided by Westat to allow for the estimation of proficiency for the same target population of students estimated by the state data. The means and standard deviations of the distributions for each scale were obtained for this sample in the same manner as described in step 1. These means and standard deviations were then linearly adjusted to reflect the reporting metric used for the national assessment (see the technical report for the NAEP 1994 national assessment.)

3) For each scale, a set of linear transformation coefficients were obtained to link the state scale to the corresponding national scale. The linking was of the form

$$Y^* = k_1 + k_2 Y$$

where

Y = a scale level in terms of the system of units of the provisional BILOG/PARSCALE scale of the Trial State Assessment scaling

Y^* = a scale level in terms of the system of units comparable to those used for reporting the 1994 national reading results

k_2 = [Standard Deviation$_{SAC}$]/[Standard Deviation$_{TSA}$]

k_1 = Mean$_{SAC}$ - k_2[Mean$_{TSA}$]

The final conversion parameters for transforming plausible values from the provisional BILOG/PARSCALE scales to the final Trial State Assessment reporting scales are given in Table 7-2. All Trial State Assessment results, including those for nonpublic schools, are reported in terms of the Y^* metric using these transformations.

7.6.5 Producing a Reading Composite Scale

For the national assessment, a composite scale was created for the fourth grade as an overall measure of reading proficiency. The composite was a weighted average of plausible values on the two content area scales (Reading for Literary Experience and Reading to Gain Information). The weights for the national content area scales were proportional to the relative importance assigned to each content area for the fourth grade in the assessment specifications developed by the Reading Objectives Panel. Consequently, the weights for each of the content areas are similar to the actual proportion of items from that content area.

The Trial State Assessment composite scale was developed using weights identical to those used to produce the composites for the 1992 and 1994 national reading assessments. The weights are given in Table 7-3. In developing the Trial State Assessment composite the weights were applied to the plausible values for each content area scale as expressed in terms of the final Trial State Assessment scales (i.e., after transformation from the provisional BILOG/PARSCALE scales.)

7.6.6 Proficiency Means for the 1994 Trial State Assessment Reading Scales

Table 7-4 shows the average reading proficiencies by scale and plausible value for fourth-grade students in the 1994 national reading public-school comparison. Average proficiences are given for each scale for each of the five plausible values and their mean. A similar table for each state is included at the beginning of each state's data codebooks.

Table 7-2
Transformation Constants for the 1994 Trial State Assessment

Scale	k_1	k_2
Reading for Literary Experience	214.64	42.15
Reading to Gain Information	210.36	42.08

Table 7-3
Weights Used for Each Scale to Form the Reading Composite

Scale	Weights
Reading for Literary Experience	.55
Reading to Gain Information	.45

Table 7-4
Average Reading Proficiences by Scale and Plausible Value
1994 National Reading Grade 4 Public-School Comparison

Scale	Data Variables	1st Value	2nd Value	3rd Value	4th Value	5th Value	Mean Value (s.e.)*
Reading for Literary Experience	RRPS11-5	213.91	214.43	214.43	214.80	214.77	214.47 (1.15)
Reading to Gain Information	RRPS21-5	209.00	209.66	209.73	209.85	210.39	209.73 (1.12)
Composite	RRPCM1-5	211.70	212.28	212.31	212.57	212.80	212.33 (1.08)

Unweighted Sample Size = 6030.0
Weighted Sample Size = 6134.1 (Sum of Variable: WEIGHT)
Estimated Population = 3162526.4 (Sum of Variable: ORIGWT)

**__Note__: The standard error is the square root of two variance components: the estimated sample variance and the variance due to measurement error.*

CHAPTER 8

CONDUCTING STATISTICAL ANALYSES WITH NAEP DATA

8.1 Introduction

Standard statistical procedures should not be applied to the NAEP Trial State Assessment data without modification because the special properties of the data affect the validity of conventional techniques of statistical inference. There are two reasons for this. First, a complex sampling scheme, rather than simple random sampling, was used to collect NAEP data. Second, because scaling models were used to summarize performance in each subject area, measurement error must be taken into account when analyzing scale-score proficiency variables.

In the NAEP sampling scheme, students do not have an equal probability of being selected. Therefore, as in all complex surveys, each student has been assigned a sampling weight. The larger the probability of selection for students within a particular demographic group, the smaller the weights for those students will be. When computing descriptive statistics or conducting inferential procedures, one should weight the data for each student. ***Performance of statistical analyses without weights can lead to misleading results.***

Another way in which the complex sample design used by NAEP differs from simple random sampling is that the NAEP sampling scheme involves the selection of clusters of students from the same school, as well as clusters of schools from urbanicity, income, and minority strata (in the case of the Trial State Assessment) and from the same geographically defined primary sampling unit, or PSU (in the case of the national assessment). As a result, observations are not independent of one another as they are in a simple random sample. Therefore, ***use of standard formulas for estimating the standard error of sample statistics such as means, proportions, or regression coefficients will result in values that are generally too small.*** The standard error, which is a measure of the variability of a sample statistic, gives an indication of how well that statistic estimates the corresponding population value. It is used to conduct tests of statistical significance. If conventional simple random sampling formulas are used to compute standard errors, too many statistically significant results will occur in most instances.

Another effect of the NAEP sampling scheme is a reduction of the effective degrees of freedom. In a simple random sample, the degrees of freedom of a variance estimate are based primarily on the number of subjects (although it also depends on the distribution of the variable under consideration). In the NAEP 1994 designs, the degrees of freedom are a function of the number of clusters of schools (for the Trial State Assessment) or clusters of PSUs (for the national assessment), rather than the number of subjects (see Chapter 4 for a discussion of the sample design). Therefore, ***the standard formulas for obtaining degrees of freedom are not valid with the NAEP data.***

Proficiencies in content areas were summarized through item response theory (IRT) scaling models, but not in the way that these models are used in standard applications in which enough responses are available from each person to estimate his or her proficiency precisely. NAEP administers relatively few items to each respondent in order to track *population* levels of proficiency more efficiently. Because the data are not intended to estimate *individual* levels of proficiency, however, more complicated analyses are required.

The following sections outline the procedures used in NAEP to account for the special properties of the data. Section 8.2 discusses the use of weights to account for the differential sampling rates and certain other adjustments, such as for nonresponse. Section 8.3 discusses jackknife procedures that can be used to estimate sampling variability. Section 8.4 describes the "plausible values" that can be used to estimate population levels of proficiency in the subject areas, and shows how to use them in

analyses. Section 8.5 suggests simpler approximations for the procedures described in 8.3 and 8.4, such as using design effects rather than the jackknife to estimate sampling variability. Although this procedure is less precise, it requires substantially less computation. We expect that the resulting degree of accuracy will be acceptable to most users of NAEP data.

8.2 Using Weights to Account for Differential Representation

The 1994 Trial State and national assessments used complex sample designs to obtain the students who were assessed. The goal of the national design was to obtain a series of samples (for the various ages and grades) from which estimates of population and subpopulation characteristics could be obtained with reasonably high precision (low sampling variability) per unit of cost. The goal of the Trial State design was to obtain a sample of students for each jurisdiction from which estimates of population and subpopulation characteristics could be obtained with approximately equal precision for all jurisdictions.

To accomplish these goals, NAEP used multistage cluster sample designs (described in Chapter 4) in which the probabilities of selection of the clusters were proportional to measures of their size. To provide improved precision in the estimation of the characteristics of various subpopulations of interest, in the national assessment some subpopulations (corresponding to students from areas with high concentrations of Black or Hispanic students and to students from nonpublic schools) were deliberately sampled at approximately twice the normal rate to obtain larger samples of respondents from those subpopulations. The result of these differential probabilities of selection for the national assessment is a series of achieved samples, each containing proportionately more members of certain subgroups than there are in the population.

Appropriate estimation of population characteristics for both the Trial State Assessment and national assessment samples must take the sampling design into account. This is accomplished by assigning a weight to each respondent, where the weight properly accounts for the sample design and, in the case of the national assessment, reflects the appropriate proportional representation of the various types of individuals in the population. These weights also include adjustments for nonresponse and, in the case of the national assessment, adjustments (known as poststratification adjustments) designed to make sample estimates of certain subpopulation totals conform to external, more accurate, estimates. An overview of the computation of these weights appears in Chapter 4. For the present purpose, it is sufficient to note that these weights should be used for all analyses, whether exploratory or confirmatory.

The 1994 Trial State Assessment database includes a number of different samples from several populations. Each of these samples has its own set of weights to be used to produce estimates about the characteristics of the population addressed by the sample (the target population). The various samples, their target populations, and their weights are discussed in the following sections.

8.2.1 The 1994 State Samples of Students

These samples, one for each jurisdiction, consist of all fourth-grade students assessed in that jurisdiction as part of the Trial State Assessment in reading. The target populations for each jurisdiction consist of all fourth-grade students enrolled in public and nonpublic schools who were deemed assessable by their school. Either of two alternatively scaled weights can be used for analyses at the student level. The first of these, ORIGWT, has been scaled so that the sum of weights for all students in each jurisdiction estimates the total number of assessable fourth-grade students in that jurisdiction's schools. The second of these, WEIGHT, is a proportional rescaling of ORIGWT, carried out so that the sum of WEIGHT across students and jurisdictions is equal to the total Trial State Assessment sample size across all jurisdictions (i.e., the total number of assessed students in the Trial State Assessment). Both weights should provide identical estimates of means, proportions, correlations, and other statistics

of interest in analyses within each jurisdiction as well as analyses involving data from more than one state.

An estimate of the proportion of students in the population who possess some characteristic can be obtained using either WEIGHT or ORIGWT as the ratio of the sum of the weights for the students with that characteristic, divided by the sum of the weights for all students sampled from that population. In the case where ORIGWT is used, the numerator of the proportion is the estimated total number of students with that characteristic and the denominator is the estimated population total. Estimated proportions can also be restricted to subpopulations. For example, the estimated proportion of all assessable students from public schools in New York is

$$\frac{WTOT(New\ York\ and\ Public)}{WTOT(New\ York)}$$

where WTOT(New York and Public) is the sum of the weights (WEIGHT or ORIGWT) of all students in New York who are in public schools and WTOT(New York) is the sum of the weights (WEIGHT or ORIGWT corresponding to the numerator) of all students in New York.

It is also clearly of interest to estimate the relative proportion of a population (say New York students) who could correctly respond to an assessment exercise. This proportion is estimated by the ratio

$$P = \frac{WTOT(New\ York,\ answered\ item\ correctly)}{WTOT(New\ York,\ presented\ the\ item)}$$

where the numerator is the sum of weights (WEIGHT or ORIGWT) of all assessed students in New York who responded to the item correctly and the denominator is the sum of weights (WEIGHT or ORIGWT corresponding to the numerator) of all students who

1) were from New York, and,

2) were presented the item (i.e., reached the item, including those who reached it and left it blank).

This total is less than WTOT(New York) because not all students are presented every item, either as a result of the spiral design or as a result of not reaching the item. However, the sample of assessed students in New York who had an opportunity to respond to the item (which includes those who did not reach the item) is itself a representative sample of the entire population of assessable students in New York.

8.2.2 Weights for Comparing Monitored and Unmonitored Sessions

In all jurisdictions that had also participated in the 1994 Trial State Assessment, one-fourth of the selected public schools were designated at random to be monitored during the assessment field period. One-half of the selected public schools were designated to be monitored in jurisdictions that had not participated in the 1994 assessment—specifically, Montana, Washington, and Department of Defense Education Activity Overseas. One-half of all nonpublic schools in every jurisdiction (regardless of 1994 participation) were designated to be monitored. Investigators may be interested in assessing the impact of monitoring on assessment performance or in otherwise comparing the samples of students in monitored and unmonitored sessions. For example, it might be of interest to compare the percentage of students from monitored sessions in New York that correctly answered a particular reading question to the corresponding percentage from unmonitored sessions. For these analyses, either WEIGHT (which sums to the overall sample size) or ORIGWT (which sums to population sizes) should be used for all analyses intended to compare statistics (such as a mean, proportion, or correlation) obtained from monitored sessions to the same statistic obtained in the unmonitored sessions. Monitor status is provided in the student file variable MONSTUD: a value of 0 indicates that the session was not monitored; a value of 1 indicates that the session was monitored.

8.2.3 The Comparison Sample from the National Assessment

One of the purposes of the Trial State Assessment was to allow each participating jurisdiction to compare its results with the nation as a whole, and with the region of the country in which that jurisdiction is located. To permit such comparisons, nationally representative samples of students were tested as part of the national assessment using the same assessment booklets as were students participating in the Trial State Assessment. The national data to which the Trial State Assessment reading results were compared came from a nationally representative sample of students in the fourth grade. This sample was a part of the full 1994 national reading assessment in which nationally representative samples of students in public and nonpublic schools from three age cohorts were assessed: students who were either in the fourth grade or 9 years old; students who were either in the eighth grade or 13 years old; and students who were either in the twelfth grade or 17 years old.

In order to allow for valid state/nation comparisons, the national comparison sample of grade-level students was created from the full national assessment sample and is included with the Trial State Assessment data files. As with the Trial State Assessment samples, two sets of weights are available for use with the national comparison sample. ORIGWT will sum to the size of the national comparison population. WEIGHT is a proportional rescaling of ORIGWT whose sum is approximately equal to the national comparison sample size. When used with standard statistical packages, both sets of weights will produce identical results for point estimates of means, proportions, standard deviations, correlations, and other such statistics.

8.2.4 School-Based Weights

The 1994 Trial State and national assessments collected questionnaire data from the assessed students' teachers about their background and instructional practices and information from administrators about aspects of the schools attended by the assessed students. Analyses of these data using the weights described above will produce results that are focused on students. For the school questionnaire data, it possible to conduct school-level analyses. The school weights SCHWTF should be used for these purposes. It should be noted that analogous teacher weights are not provided and the NAEP samples were not selected to contain representative samples of teachers. Analyses of the teacher questionnaire data should be restricted to student-level analyses.

8.3 Procedures Used by NAEP to Estimate Sampling Variability (Jackknifing)

This section describes how the sampling variability of statistics based on the NAEP data can be estimated. The jackknife variance estimator described below gives fairly precise estimates of the total sampling error for population estimates derived from NAEP student and school data, and for conducting multivariate analyses. To aid secondary users who have fewer resources than those available for the NAEP reports, Section 8.5 provides a less expensive approximation for estimating sampling variances.

A major source of uncertainty in the estimation of the value in the population of a variable of interest exists because information about the variable is obtained on only a sample from the population. To reflect this fact, it is important to attach to any statistic (e.g., a mean) an estimate of the sampling variability to be expected for that statistic.

Estimates of sampling variability provide information about how much the value of a given statistic would be likely to change if the statistic had been based on another equivalent sample of individuals drawn in exactly the same manner as the achieved sample. Consequently, the estimation of the sampling variability of any statistic must take into account the sample design.

The NAEP samples are obtained via a stratified multistage probability sampling design that includes, in the case of the national comparison samples, provisions for sampling certain subpopulations at higher rates. Additional characteristics of the samples include adjustments for both nonresponse and, for the national comparison samples, poststratification. The resulting samples have different statistical characteristics than those of a simple random sample. In particular, because of the effects of cluster selection (students within schools) and nonresponse and other weighting adjustments, observations made on different students cannot be assumed to be independent of each other. Furthermore, to account for the differential probabilities of selection and the various sample weighting adjustments, each student has an associated sampling weight that must be used in the computation of any statistic and is itself subject to sampling variability.

Treatment of the data as a simple random sample, with disregard for the special characteristics of the NAEP sample design, will produce underestimates of the true sampling variability. A procedure known as jackknifing is suitable for estimating sampling errors from such a complex design. This procedure has a number of properties that make it particularly suited to the analysis of NAEP data:

1) It provides unbiased estimates of the sampling error arising from the complex sample selection procedure for linear estimates such as simple totals and means, and does so approximately for more complex estimates.

2) It reflects the component of sampling error introduced by the use of weighting factors, such as nonresponse adjustments, that are dependent on the sample data actually obtained.

3) It can be adapted readily to the estimation of sampling errors for parameters estimated using statistical modeling procedures, as well as for tabulation estimates such as totals and means.

4) Once appropriate weights are derived and attached to each record, jackknifing can be used to estimate sampling errors. A single set of replicate weights is required for all tabulations and model parameter estimates that may be needed.

The jacknife process of replication involves repeatedly selecting portions of the sample (replicates) and calculating the desired statistic (replicate estimates). The variability among the calculated replicate estimates is then used to obtain the variance of the full-sample estimate.

In each jurisdiction, replicates were formed in two steps. First, each school was assigned to one of a maximum of 62 replicate groups, each group containing at least one school. In the next step, a random subset of schools (or, in some cases, students within schools) in each replicate group was excluded. The remaining subset and all schools in the other replicate groups then constituted one of the 62 replicates.

Replicate groups were formed separately for public and nonpublic schools. Once replicate groups were formed for all schools, students were then assigned to their respective school replicate groups.

Public Schools. These schools were sorted according to the jurisdiction, monitoring status, and, within monitoring status, the order in which they were selected from the sampling frame. The schools were then were grouped in pairs or, occasionally, triples. The pairing was done such that no single pair contained schools with different monitoring status.

For the largest schools selected with certainty, the replicate groups consisted of random subgroups of students within each certainty school.

The purpose of this scheme was to assign as many replicates to a jurisdiction's public schools as permitted by the design, to a maximum of 62. When more than 62 replicates were assigned, the procedure ensured that no subset of the replicate groups (pairs of noncertainty schools, individual certainty schools, or groups of these) was substantially larger than the other replicate groups. The aim was to maximize the degrees of freedom

available for estimating variances for public-school data.

A single replicate was formed by dropping one member of a given pair. This process was repeated successively across pairs, giving up to 62 replicates.

Nonpublic Schools. Replicate groups for noncertainty nonpublic schools were formed in one of the two methods described below. If any of the following conditions was true for a given jurisdiction, then the subsequent steps were taken to form replicate groups. Here, the numbering started at 62 down to the last needed number.

Conditions for Method 1:

- fewer than 11 nonpublic noncertainty schools;
- fewer than 2 Catholic noncertainty schools; or
- fewer than 2 nonCatholic noncertainty schools.

Steps for Method 1:

- all schools were grouped into a single replicate group;
- schools were randomly sorted; and
- starting with the second school, replicates were formed by consecutively leaving out one of the remaining *n - 1* schools; each replicate included the first school.

When a given jurisdiction did not match conditions of the first method (i.e., when all of the following conditions were true) then the preceding steps were repeated separately for two replicate groups, one consisting of Catholic schools and one consisting of nonCatholic schools.

Conditions for Method 2:

- more than 10 nonpublic noncertainty schools;
- more than 1 Catholic noncertainty school; and
- more than 1 nonCatholic noncertainty school.

For jurisdictions with certainty nonpublic schools (Delaware, District of Columbia, and Hawaii) each school was assigned to a single group. Prior to this assignment, schools were sorted in descending order of the estimated grade enrollment. The group numbering started at the last number where the noncertainty nonpublic schools ended. A replicate was formed by randomly deleting one half of the students in a certain school from the sample. This was repeated for each certainty school.

Again, the aim was to maximize the number of degrees of freedom for estimating sampling errors for nonpublic schools (and indeed for public and nonpublic schools combined) within the constraint of forming 62 replicate groups. Where a jurisdiction had a significant contribution from both Catholic and nonCatholic schools, we ensured that the sampling error estimates reflected the stratification on this characteristic.

These pairings are identified by the variables JKPAIR and JKREP2 on the national comparison sample student data files and REPGRP1 and REPGRP2 on the state student data files; membership within the pair (or triple) is identified by the variable JKUNIT on the national comparison sample student data files and DROPGRP on the state student data files (corresponding replicate variables for the school samples exist on the school files).

Components of the sampling variability of an estimate are each estimated as the squared difference between the value of the statistic for the complete sample and a pseudoreplicate formed by recomputing the statistic on a specially constructed pseudodataset. This pseudodataset is created from the original dataset by eliminating one member of a pair and replacing it with a copy of the remaining unit or units in the pair. For computational purposes, a pseudoreplicate associated with a given pair is the original dataset with a different set of weights (referred to as the student replicate weights SRWT01 through SRWT62 on the data files, where SRWTi is for the i^{th} pair). This set of weights allows measurement of the total effect of replacing one member of the pair with a copy of the other(s), including adjustments for nonresponse and, for the national comparison sample, poststratification. The

i^{th} pseudoreplicate for a given statistic is obtained by recalculating the statistic using the weights SRWTi instead of the original sampling weights.

As a specific example of the use of the student replicate weights, let $t(\underline{y},\underline{w})$ be any statistic that is a function of the sample responses $\underline{y}$ and the weights $\underline{w}$ that estimates population value T. For example, t could be a weighted mean, a weighted percent-correct point, or a weighted regression coefficient. The $t(\underline{y},\underline{w})$, computed with the sampling weights (ORIGWT on the data files) is the appropriate sample estimate of T. To estimate $\hat{Var}(t)$, the sampling variance for this statistic, proceed in the following manner:

1) For each of the 62 pairs of first-stage units, compute the associated pseudoreplicate for the statistic. For the i^{th} pair, this is

$$t_i = t(\underline{y}, \underline{SRWTi}) \ ,$$

which is the statistic t recalculated by using SRWTi instead of the original sampling weights.

2) The estimated sample variance of t is

$$\hat{Var}(t) = \sum_{i=1}^{62} (t_i - t)^2 \quad .$$

We refer to this estimation technique as the multiweight jackknife approach. Tables 10-7 and 10-8 in Chapter 10 provide SPSS-X and SAS code for carrying out the above in the special case of a weighted mean.

Replicate weights have been provided for:

1)	Overall analyses in each state in the Trial State Assessment samples	SRWT01 to SRWT62
2)	For monitored/unmonitored comparisons within each state in the Trial State Assessment sample	SRWT01 to SRWT62
3)	For school-based analyses in each state for the Trial State Assessment samples	SCHWT01 to SCHWT62
4)	Overall analyses in the national comparison sample	SRWT01 to SRWT62
5)	For school-based analyses in the national comparison sample	SCHWT01 to SCHWT62

In addition, for analyses comparing national and state results, or for comparisons among jurisdictions, an appropriate single set of replicate weights can be formed for the merged dataset by using the relevant set of replicate weights for each given component. That is, the first replicate estimate of a difference between a national student-level estimate and that for a given jurisdiction is obtained by using the replicate weight SRWT01 for each record in the national sample and for each record in the particular state sample, and calculating the difference between the respective replicated national and state estimates.

As a very simple example of how the jackknife variance estimate is computed, consider the following cut-down example, designed to demonstrate the steps. Although the full set of NAEP data consists of thousands of observations and 62 student replicate weights, for the example we will consider a dataset (Table 8-1) with eight observations and two student replicate weights. Furthermore, the weights have been simplified for clarity.

Table 8-1
Example Dataset to Demonstrate the Jackknife

First-Stage Unit	JKPAIR	JKUNIT	Y	ORIGWT	SRWT01	SRWT02
1	1	1	5	10	20	10
1	1	1	4	9	18	9
2	1	2	6	12	0	12
2	1	2	3	8	0	8
3	2	1	8	4	4	8
3	2	1	9	6	6	12
4	2	2	7	5	5	0
4	2	2	10	4	4	0

In the example dataset there are four first-stage units, 1 through 4, each consisting of two of the eight observations. The first-stage units are divided into two pairs, as identified by the column JKPAIR. Within each of those pairs, one first-stage unit is designated as the first member of the pair (JKUNIT = 1) while the other is designated as the second (JKUNIT = 2). The statistic of interest is the weighted average of the response Y using the weights ORIGWT, and is equal to

$$t = NUM/DEN = 5.914$$

where

$$NUM = 10\times5 + 9\times4 + 12\times6 + 8\times3 + 4\times8 + 6\times9 + 5\times7 + 4\times10 = 343$$

is the weighted sum of the responses and

$$DEN = 10 + 9 + 12 + 8 + 4 + 6 + 5 + 4 = 58$$

is the sum of the weights ORIGWT.

The first pseudoreplicate of the statistic t is the weighted mean recomputed using the SRWT01 as the weights and is

$$t_1 = NUM_1/DEN_1 = 5.842$$

where

$$NUM_1 = 20\times5 + 18\times4 + 0\times6 + 0\times3 + 4\times8 + 6\times9 + 5\times7 + 4\times10 = 333$$

and

$$DEN_1 = 20 + 18 + 0 + 0 + 4 + 6 + 5 + 4 = 57.$$

Similarly, $t_2 = 354/59 = 6$ is the weighted mean computed using SRWT02 as the weights. The jackknife variance estimate is then

$$\hat{Var}(t) = \sum_{i=1}^{2} (t_i - t)^2 = (t_1 - t)^2 + (t_2 - t)^2 = (-0.072)^2 + (0.086)^2 = .01258$$

and the jackknife standard error of t is .112, the square root of the variance.

8.3.1 Degrees of Freedom of the Jackknife Variance Estimate

The effective number of degrees of freedom of the variance estimate $\hat{Var}(t)$ will be at most equal to the number of pairs used in forming the pseudoreplicates. The number of degrees of freedom in sampling from normally distributed variates with uniform variances is sufficient

information to indicate the variability of the variance estimate, and is equal to the number of independent pieces of information used to generate the variance. For each assessment sample, the pieces of information are the 62 squared differences $(t_i - t)^2$, each supplying at most one degree of freedom, regardless of how many individuals were sampled within any replicate groups.

The effective number of degrees of freedom of the sample variance estimator can be less than the number of pairs (62) if the differences are not normally distributed or if some of the squared differences $(t_i - t)^2$ are markedly different in magnitude than others. An extreme case of the latter is when one or more of the t_i are identical to t, so that $(t_i - t)^2 = 0$. This may happen, for example, when the statistic t is a mean for a subgroup, such as a type of location, and no members of that subgroup come from the pair i. Such a pair contributes zero to the effective number of degrees of freedom of the variance estimate.

An estimate of the effective number of degrees of freedom for $\hat{Var}(t)$ comes from an approximation due to Satterthwaite (1941). (See Cochran, 1977, p. 96, for a discussion.)

If the t_i are normally distributed, the effective number of degrees of freedom using this approximation is

$$df_{eff} = \frac{[\sum_{i=1}^{K} (t_i - t)^2]^2}{\sum_{i=1}^{K} (t_i - t)^4},$$

where K is the number of pairs used (for the Trial State Assessment, K = 62).

8.3.2 Estimation of Subpopulations with Appropriate Jackknife Standard Errors

As stated in Section 8.2.1, the variable WEIGHT on the student files is a proportional rescaling of the variable ORIGWT. The factor used to calculate the rescaled weight (WEIGHT) from the original weight (ORIGWT) for the state samples was 23.3695. The corresponding factor for the national comparison sample was 515.5679. The excluded student weights (XWEIGHT) and the school weights (SWEIGHT) remain in the original metric; there are no rescaled weights for these samples.

These factors are required to estimate the number in a population and compute the corresponding jackknife standard error, which estimates how well the number in the population has been estimated. The replicate weights SRWT01 to SRWT62 are in the ORIGWT metric. To use the jackknife procedure with WEIGHT, multiply each replicate weight by the appropriate factor, yielding new replicate weights to be used in the jackknife procedure. The resulting standard error will be the appropriate estimate of the variability of the weights.

8.4 Procedures Used by NAEP to Handle Imprecision of Individual Measurement

Jackknifing provides a reasonable estimate of uncertainty due to the sampling of respondents when the variable of interest is observed without error from every respondent. Population percents correct for cognitive items meet this requirement, but scale-score proficiency values do not. The item response theory (IRT) models used to summarize performance in a subject area or subarea posit an unobservable proficiency variable θ to summarize performance on the items in that area. The fact that θ values are not observed even for the respondents in the sample requires additional statistical machinery to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. To this end, we have adapted Rubin's (1987) "multiple imputations" procedures for missing data to the context of latent variable models to produce the "plausible values" that appear in the NAEP 1994 secondary-use data files.

The essential idea of plausible values methodology is that even though we do not observe the θ value of respondent i, we do observe other kinds of variables that are related to it: x_i, the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i, the respondent's answers to demographic and background variables. Suppose we would like to draw inferences about a number $T(\Theta,Y)$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\theta,y)$, where $(\theta,y) \equiv (\theta_1,y_1,...,\theta_N,y_N)$, and that we could estimate the variance in t around T due to sampling respondents by the function $U(\theta,y)$. Given that observations consist of (x_i,y_i) rather than (θ_i,y_i), we can approximate t by its expected value conditional on (x,y), or

$$t^*(x,y) = E[t(\theta,y) \mid x,y] = \int t(\theta,y)p(\theta \mid x,y)\, d\theta \quad .$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\theta_i \mid x_i,y_i)$, which are obtained for all respondents by the method described in Chapter 7. Let $\hat{\theta}_m$ be the m^{th} such vector of "plausible values." It is a plausible representation of what the true θ might have been, had we been able to observe it. The following steps describe how an estimate of a scalar statistic $t(\theta,y)$ and its sampling variance can be obtained from M (>1) such sets of plausible values. (Note: five sets are provided on the data files for each subject area or subarea analyzed by these procedures.)

1) Using each set of plausible values $\hat{\theta}_m$ in turn, evaluate t as if the plausible values were true values of θ. Denote the results $\hat{t}_m$, for m=1,...,M.

2) Using the multiple weight jackknife approach, compute the estimated sampling variance of $\hat{t}_m$, denoting the result as U_m.

3) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \hat{t}_m / M \quad .$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} U_m / M \quad .$$

5) Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} (\hat{t}_m - t^*)^2 / (M-1) \quad .$$

6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\, B_M \quad .$$

***Note**: NAEP reports use a single jackknife estimate U_m in place of the average of five, as would be required for U^*; see Section 8.5.*

Suppose that the statistic $[t(\theta,y) - T]/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the distribution of $(t^* - T)/V^{1/2}$ is also approximately t, with degrees of freedom given by

$$v = (M - 1)(1 + r_M^{-1}) \frac{d}{d + r_M^{-2}(M-1)}$$

where r_M is the relative increase in variance due to not observing θ values:

$$r_M = (1 + M^{-1})\, B_M / U^* \quad .$$

When B is small relative to U, and d is large, a normal approximation suffices. This is the case with main NAEP reporting variables, and the normal

approximation is routinely applied to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_m and U^* are covariance matrices, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity

$$(T - t^*)\ V^{-1}\ (T - t^*)`$$

is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of r_M:

$$r_M = (1 + M^{-1})\ Trace(B_M U^{*-1}) / k \quad .$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

Computation of statistics t^* involving the plausible values and categories of variables included in the conditioning variables y (described in Chapter 7) yields consistent estimates of the corresponding population values T. *Statistics involving background variables y that were <u>not</u> conditioned on are subject to biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variable(s) to the variables that were conditioned on. The <u>direction</u> of the bias is typically to underestimate the effect of nonconditioned variables.*[6] For a given statistic t^* involving one or more nonconditioned background variables, the <u>magnitude</u> of the bias is related to the fraction of information about T that is missing because θ is not observed:

$$\gamma_M = \frac{r_M + 2/(\nu + 3)}{r_M + 1} \quad .$$

[6]For details, see Section 10.3.5 of *Implementing the New Design: The NAEP 1983-84 Technical Report,* Section 8.4.3 of *Expanding the New Design: The NAEP 1985-86 Technical Report,* and Mislevy, 1991.

8.5 Approximations

A jackknife estimate of the variability of a statistic based on one or more observed NAEP variables in the 1994 sample requires computing the statistic 63 times. Estimating the variability for a statistic involving a scale-score could require computing the statistic as many as 315 times, including 53 runs to obtain a variance estimate for each of five sets of plausible values. Because the cost of the full procedure may well prove prohibitive in many studies, approximate procedures that produce reasonable estimates at lower costs are provided below. Section 8.5.1 gives approximations for sampling variation; 8.5.2 gives approximations for variation due to measurement error associated with scale-scores; 8.5.3 discusses strategies for combining the suggestions in 8.5.1 and 8.5.2.

8.5.1 Approximations for Sampling Variability

The major computational load in calculating uncertainty measures for any statistic exists in the computation of the uncertainty due to sampling variability. As noted in the last section, a jackknife estimate of the variability of a statistic based on one or more observed NAEP variables in the 1994 main assessment samples requires computing the statistic 63 times. This section describes a less computationally intensive approximation to sampling variability of any statistic.

As indicated in Section 8.3, it is inappropriate to estimate the sampling variability of any statistic based on the NAEP database by using simple random sampling formulas. These formulas, which are the ones used by most standard statistical software such as SPSS and SAS, will produce variance estimates that are generally much smaller than is warranted by the sample design.

It may be possible to account approximately for the effects of the sample design by using an inflation factor, the design effect, developed by Kish (1965) and extended by Kish and Frankel (1974). The design effect for a statistic is the ratio of the actual

variance of the statistic (taking the sample design into account) over the simple random sampling variance estimate based on the same number of elements. The design effect may be used to adjust error estimates based on simple random sampling assumptions to account approximately for the effect of the design. In practice, this is often accomplished by dividing the total sample size by the design effect and using this effective sample size in the computation of errors. Note that the value of the design effect depends on the type of statistic computed and the variables considered in a particular analysis as well as the clustering effects occurring among sampled elements and the effects of any variable weights resulting from variable overall sampling fractions.

On the basis of empirical results and theoretic considerations, Kish and Frankel (1974) have developed several conjectures about design effects:

1) Generally, the design effects for complex statistics from complex samples are greater than 1, causing variances based on simple random sampling assumptions to tend to be underestimated.

2) The design effects for regression coefficients tend to be smaller than the corresponding design effects for means of the same variables. Hence, these latter estimates, which are more easily computed, tend to overestimate the design effects of complex statistics. For correlation coefficients and partial correlation coefficients, the design effect for the mean should be used (Skinner, Holt, & Smith, 1989, p. 70).

3) The size of the design effects of complex statistics tends to parallel those of means; variables with a high design effect of the mean also tend to have high design effects for complex statistics involving those variables.

To incorporate the design effect idea in a statistical analysis, proceed in the following manner:

1) For a given class of statistics (e.g., means, proportions, regression coefficients), compute the jackknife variance described in Section 8.3.1 for a number of cases. The cases should cover the range of situations for which the approximation is to be used. If various subpopulations are to be considered, it is important to have information on the relative variability within each subgroup. This is especially important if certain subgroups are more highly clustered in the sample.

2) For the identical cases, compute the simple random sampling variance given the elements in the sample. To account properly for the difference between the number of individuals being sampled and the total of the sampling weights, the weights should be scaled so that their sum equals the sample size.

3) For each case, compute the design effect where the design effect for case j is

$$deff_j = Var_{JK}(t_j)/Var_{CON}(t_j) \quad ,$$

the ratio of the jackknife variance estimate of the statistic to its simple random sampling variance estimate.

4) If the design effects for the various cases are tolerably similar, choose an overall composite design effect. If the design effects for certain subgroups appear to cluster around a markedly different value from the remaining cases, treat those subgroups separately.

5) In the case that a consistent overall design effect has been found:

   a) Rescale the weight of each individual so that the sum of the scaled weights is equal to the effective sample size

$$N_{eff} = \frac{sample\ size}{design\ effect}$$

(that is, multiply each weight by N_{eff}/W_{TOT}, where W_{TOT} is the sum of the original weights).

b) Conduct a traditional weighted analysis using these scaled weights.

8.5.2 Approximations for Measurement Error Variability

A second method of reducing costs applies to statistics that involve scale-score proficiency values: using fewer runs on plausible value sets. A statistic t^* based on a single set of plausible values has the same expectation as the average of five, but cannot take into account the uncertainty caused by the fact that θ is unobserved. Compared to using all five sets of plausible values, using at least two but fewer than five sets to evaluate a statistic allows one to account for this component of uncertainty and reduce costs at the same time. One merely applies the formulas given in Section 8.4 with M=2, 3, or 4, as appropriate. It may be seen that the resulting decrease in computation is accompanied by an increase in total variance associated with t^*, but one that may be worth the price.

***Note**: It is not recommended to compute the average of the five plausible values associated with each respondent, then analyze these averages. This procedure does not generally give the correct value of a statistic.*

8.5.3 Approximating Both Sampling and Measurement Variability

Full implementation of the NAEP procedures for estimating the variability of a statistic involving a scale-score variable requires 315 runs. Combining the approximations suggested above in various ways allows the researcher to trade off precision and cost in a manner that suits his or her needs. Some options are discussed below.

1) *Full implementation* (315 runs). This option is the most costly, but the most precise. Each estimate of a statistic and each jackknife estimate of its sampling variance is calculated on all five sets of plausible values.

2) *Estimate based on five sets of plausible values, jackknife based on one set of plausible values* (67 runs). This option involves computing the statistic t^* exactly as described in Section 8.4, but basing its variance estimate on the sum of B (the variance of five $\hat{t}_m$ estimates) and one U_m value (rather than the average of five). This is the option routinely used by NAEP in its own reports. It gives the same point estimate of T as the full implementation, but the variance estimate, while still consistent, is less precise. Using the jackknife as opposed to a design effect accounts for 62 of the runs, but allows for differential impacts of the respondent-sampling design upon the variability of different statistics.

3) *Estimate based on five sets of plausible values, design effect for sampling variance* (5 runs, assuming a design effect has already been estimated). This option gives the same point estimate of T as options 1) and 2), but a less precise estimate of its variability. It is obtained by computing t^* and $V(t^*)$ just as described in Section 8.4, but with each U_m value obtained by boosting the SRS sampling variance estimates in accordance with a design effect as described in Section 8.5.1. Note that additional initial runs will be needed to estimate the design effect.

4) *Estimate based on M sets of plausible values, where 1<M<5; design effect for sampling variance* (M runs). The point estimates provided by this option now differ from those in previous options. They have the same expectations as those described above, but now the point estimates themselves, rather than just the estimates of their variability, are less precise. By using at least two sets of plausible values, however, the researcher ensures that both the sampling and the measurement components of variability are taken into

account. This option is attractive for researchers who have very limited resources.

5) *Estimate based on one set of plausible values, design effect for sampling variance* (1 run). The point estimate obtained here has the same expected value as those described above, but is again less precise. Measurement variability cannot be estimated with only one set of plausible values, and statements of variability or significance tests based on sampling variability only are incorrect because they underestimate variability. The degree of underestimation depends on the statistic being computed. For population or subpopulation averages of proficiency on background variables included in all booklets, the degree of underestimation of variability is roughly 20 percent (Rubin, 1987, Table 4.1). For statistics that are more complex or involve background variables that appear on only a subset of booklets, the underestimation can easily exceed 50 percent. This option is not recommended for such statistics.

Again, a strategy that *should not* be considered deserves repeated emphasis: *Computing the average of the five plausible values associated with each respondent, then analyzing these averages, does* ***not*** *generally give the correct value of a statistic.*

8.6 Additional Sources of Error

In addition to errors due to sampling and imprecision of individual measurement, NAEP results are also subject to other kinds of errors, including the effects of necessarily imperfect adjustment for student and school nonresponse and other largely unknowable effects associated with the particular instrumentation and data collection methods used. Nonsampling errors can be attributed to a number of sources: inability to obtain complete information about all selected students in all selected schools in the sample (some students or schools refused to participate, or students participated but answered only certain items); ambiguous definitions; differences in interpreting questions; inability or unwillingness to give correct information; mistakes in recording, coding, or scoring data; and other errors of collecting, processing, sampling, and estimating missing data. The extent of nonsampling errors is difficult to estimate. By their nature, the impacts of such errors cannot be reflected in the data-based estimates of uncertainty.

Users of NAEP data should also be aware that there are additional components of variance, due to the statistical nature of the scaling and linking process, that are not included in the various estimation procedures discussed in Sections 8.3 and 8.4. In NAEP, as in other applications of IRT, item parameters are unknown; estimates must be used. Research is underway on how uncertainty associated with item parameter estimates affects the estimation of proficiency distributions (see, e.g., Tsutakawa & Johnson, 1990). The estimation error associated with scale linking (such as the linking of the Trial State Assessment results to the national results, described in Section 7.5.4) represents another source of uncertainty. Some preliminary investigations into estimating the uncertainty associated with scale linking have been carried out by Sheehan and Mislevy (1988) and Johnson, Mislevy, and Zwick (1990). At present, standard errors for NAEP results reflect only the estimation due to sampling of students and due to imprecision of individual measurement. Research is underway to determine mechanisms for including other sources of uncertainty into the variance estimation procedures.

8.7 A Note Concerning Multiple Comparisons

If many statistical tests are conducted at one time, it is likely that significance tests will overstate the degree of statistical significance of the results. In the preceding sections, we noted that because of the design of the NAEP sample, conventional significance tests will overstate significance, because they fail to consider the effects of clustering. In contrast, the problem of multiple comparisons noted here is independent of sample design; it arises even if one uses the appropriate statistical tests described previously. The problem arises because the more statistical tests are

calculated, the more likely it becomes that one will find a "significant" finding because of chance variation. In other words, the chance of a type I error—a spurious "significant" finding—rises with the number of tests conducted.

More technically, if J multiple hypothesis tests are performed, each with a type I error rate (the probability of rejecting the null hypothesis when the null hypothesis is true) of α, the type I error rate for the entire set of tests could be as high as $J\alpha$. Therefore, it is desirable to use a multiple comparison procedure to control the overall error rate for the entire set of hypothesis tests. In the present case, it is advantageous to use a procedure that allows control of the error rate for sets of varying size that may include both pairwise and complex comparisons. (An example of a complex contrast is a comparison of one group to the average of two other groups.) The Dunn-Bonferroni approach is, therefore, a good choice. To apply this method in its simplest form, we need only decide at what level we wish to control the setwise error rate (α_S) and then set the type I error rate for each comparison equal to $\alpha_C = \alpha_S / J$, where J is the number of comparisons.

For example, suppose we wanted to perform three pairwise comparisons between regional groups, as well as one complex comparison, controlling α_S at .05. The type I error rate for each comparison should be set at $\alpha_C = \alpha_S / J = .05/4 = .0125$. The required critical value can be obtained from a table of the Bonferroni t-statistic (Miller, 1981, p. 238) with the appropriate degrees of freedom.

CHAPTER 9

CONTENT AND FORMAT OF DATA FILES, LAYOUTS, AND CODEBOOKS

9.1 Introduction

This chapter describes in detail the content and format of each secondary-use data file and the accompanying printed layouts and codebooks. Each data package contains a data file for each student sample and questionnaire instrument. Three other types of files are provided for each data file: a set of SPSS control statements for generating an SPSS system file; a set of SAS control statements for generating a SAS system file; and a machine-readable catalog file containing parameter data and information for each field in a data record.

The accompanying printed documentation contains a file layout and data codebook for each data file. Each layout contains the essential processing and labeling information on one line for each data field. Each codebook contains more descriptive information for each field.

9.2 Data Files

There are four file types for each sample administered in the Trial State Assessment. Files are arranged by sample within file type. The order, names, and other characteristics of the files are given in Tables 9-1 and 9-2. The files distributed on CD-ROM follow the DOS naming convention of an eight-character file name and a three-character file type, separated by a period. The files distributed on IBM tapes have similar names prefixed with "NAE." The files are named according to the following convention:

The first index level (up to the first period) designates the sample:

TSR1STUD	1994 state sample, grade 4 student data
TSR1EXCL	1994 state sample, grade 4 excluded student data
TSR1SCHL	1994 state sample, grade 4 school data
NCR1STUD	1994 national comparison reading sample, grade 4 student data
NRS1SCHL	1994 national comparison reading sample, grade 4 school data

The second index level is the file type:

DAT	The raw data file
SAS	The SAS control statements for generating a SAS system file
SPS	The SPSS control statements for generating an SPSS system file
CAT	A machine-readable catalog of item and variable information

9.2.1 Respondent Data

Depending on the sample, each raw data file contains one record per student, excluded student, or school. All raw data files are rectangular—that is, record lengths are fixed and a given variable always occurs in the same position on every record within a file. The NAEP data files are structured to facilitate matching among the three samples (student, excluded student, and school). The teacher data have already been linked with the appropriate students on the state and national comparison samples. For the purposes of analysis and reporting, only two types of linkages are valid:

1) school with student and teacher (state and national)

2) school with excluded student (state only)

Table 9-1
NAEP 1994 State Reading Data Package Description: Grade 4

Files	Record Length	Number of Records	File Name**
Data Files			
1. State Reading Student Sample	1481	*	TSR1ST*xx*.DAT
2. State Reading Excluded Student Sample	651	*	TSR1EX*xx*.DAT
3. State Reading School Sample	711	*	TSR1SC*xx*.DAT
4. National Reading Comparison Student Sample	1524	7382	NCR1STUD.DAT
5. National Reading Comparison School Sample	846	293	NCR1SCHL.DAT
SPSS Control Statement Files			
6. State Reading Student Sample	80	1844	TSR1STUD.SPS
7. State Reading Excluded Student Sample	80	409	TSR1EXCL.SPS
8. State Reading School Sample	80	478	TSR1SCHL.SPS
9. National Reading Comparison Student Sample	80	1858	NCR1STUD.SPS
10. National Reading Comparison School Sample	80	549	NCR1SCHL.SPS
SAS Control Statement Files			
11. State Reading Student Sample	80	1073	TSR1STUD.SAS
12. State Reading Excluded Student Sample	80	290	TSR1EXCL.SAS
13. State Reading School Sample	80	331	TSR1SCHL.SAS
14. National Reading Comparison Student Sample	80	1094	NCR1STUD.SAS
15. National Reading Comparison School Sample	80	388	NCR1SCHL.SAS
16. Format Library Generator	80	438	TSR1.FMT
Machine-Readable Catalog Files			
17. State Reading Student Sample	1402	618	TSR1STUD.CAT
18. State Reading Excluded Student Sample	1402	177	TSR1EXCL.CAT
19. State Reading School Sample	1402	212	TSR1SCHL.CAT
20. National Reading Comparison Student Sample	1402	632	NCR1STUD.CAT
21. National Reading Comparison School Sample	1402	251	NCR1SCHL.CAT

**Note*: *Number of records varies by jurisdiction; see Table 9-2 for record counts for each jurisdiction.*
***Note*: *The use of "xx" in a file name refers to the jurisdiction code (i.e., NJ, PA, etc.). Jurisdiction codes are included in Table 9-2 and are used in place of "xx."*

Table 9-2
NAEP 1994 Reading File Record Counts by Jurisdiction

Jurisdiction		Grade 4 Data File Record Counts		
		Student	Excluded Student	School
AL	Alabama	2845	163	108
AZ	Arizona	2651	191	104
AR	Arkansas	2689	167	104
CA	California	2401	358	102
CO	Colorado	2860	204	116
CT	Connecticut	2868	237	113
DE	Delaware	2783	146	73
DD	Dept of Defense	2413	108	81
FL	Florida	2933	302	118
GA	Georgia	2982	164	113
GU	Guam	2575	192	30
HI	Hawaii	3147	141	123
ID	Idaho	2692	141	105
IN	Indiana	2874	153	109
IA	Iowa	3086	133	123
KY	Kentucky	3036	108	113
LA	Louisiana	3170	165	122
ME	Maine	2521	257	112
MD	Maryland	2830	205	111
MA	Massachusetts	2819	237	114
MI	Michigan	2142	122	82
MN	Minnesota	3045	115	120
MS	Mississippi	2918	168	110
MO	Missouri	3042	152	124
MT	Montana	2649	86	118
NE	Nebraska	2606	103	120
NH	New Hampshire	2197	132	86
NJ	New Jersey	2888	155	113
NM	New Mexico	2826	239	114
NY	New York	2864	221	111
NC	North Carolina	2833	169	105
ND	North Dakota	2797	65	131
PA	Pennsylvania	2717	137	105
RI	Rhode Island	2696	133	109
SC	South Carolina	2863	185	109
TN	Tennessee	1998	112	76
TX	Texas	2454	288	97
UT	Utah	2733	138	105
VA	Virginia	2870	214	112
WA	Washington	2737	143	104
WV	West Virginia	2887	212	118
WI	Wisconsin	2719	181	110
WY	Wyoming	2699	116	112
Total		**118355**	**7358**	**4585**

The principal linkage between files is accomplished through the scrambled primary sampling unit/school code field, which is named SCRPSU on the student data files, SSCRPSU on the school data files, and XSCRPSU on the excluded student data files. All files are sorted by this field to permit direct match-merging without the need to re-sort.

Because of the nature of the PBIB spiral design (see Chapter 3), students were administered different blocks of items. As a result, each student record contains blank spaces for the item blocks that were not included in the student's assessment booklet (missing by design). Fields are also blank for items that did not appear in booklets because of a printing error (e.g., incorrect block in booklet, missing pages) and for the professionally scored items that were not included in reliability checks (see Section 5.4 in Chapter 5). Additionally, items that were either missed by the scorers or given erroneous codes by the scorers were coded as blank fields.

Special codes (Table 9-3) were assigned to responses that were:

- illegible or illiterate, erased or crossed out;
- off task or "I don't know";
- omitted;
- not reached; and
- multiple responses.

9.2.2 SPSS and SAS Control Statement Files

All respondent data files in the data package are accompanied by separate control files to facilitate the creation of SPSS and SAS system files. These control files include statements for variable definitions, variable labels, missing value codes, value labels, and an optional section for creating and storing scored variables. Each set of control statements also generates unweighted descriptive statistics of the reporting variables for the related data file and a listing of the contents of the saved system file.

Specific details on the structure and use of these control files are provided in Chapter 10.

9.2.3 Machine-Readable Catalog Files

The machine-readable catalog files are designed primarily for users who want to use a programming language or package other than SAS or SPSS to analyze the data. These files may also be processed by SAS or SPSS to produce listings or informational reports.

Each catalog file contains a record for each variable or item on its corresponding data file. Table 9-4 contains the machine-readable catalog data layout. Specific information concerning the contents of the catalogs is provided on the following page.

Table 9-3
Special Response Codes

Code (Width = 1)	Code (Width = 2)	Definition
5	55	ILLEGIBLE/ILLITERATE, ERASED, or CROSSED OUT (constructed-response items)
7	77	"I DON'T KNOW" (multiple-choice items) "I DON'T KNOW" or OFF TASK (constructed-response items)
8	88	OMITTED
9	99	NOT REACHED
0	0	MULTIPLE RESPONSE (multiple-choice items)

Table 9-4
NAEP 1994 State Machine-Readable Catalog File Layout

Start and End Columns	Field Width	Field Type	Field Description	Comments
1 - 4	4	N	Field Sequence Number	
5 - 12	8	A	Field Name	NAEP Ident.
13 - 16	4	N	Start Column	
17 - 20	4	N	End Column	
21 - 22	2	N	Field Width	
23 - 23	1	N	Decimal Places	
24 - 24	1	N	Field Type	
25 - 27	3	N	Minimum Valid Response	
28 - 30	3	N	Maximum Valid Response	
31 - 32	2	N	Minimum Correct Response	
33 - 34	2	N	Maximum Correct Response	
35 - 36	2	N	Illegible/Illiterate Code	
37 - 38	2	N	Nonrateable Response Code	
39 - 40	2	N	I Don’t Know Response Code	
41 - 42	2	N	Omit Code	
43 - 44	2	N	Not Reached Code	
45 - 46	2	N	Multiple Response Code	
47 - 96	50	A	Name/Description	
97 - 104	8	A	Alternate NAEP Identification	
105 - 105	1	N	Scaling Category	
106 - 106	1	N	Number of Response Categories	
107 - 146	40	N	IRT Parameters	(5F8.5) Format
147 - 148	2	N	Number of Data Codes and Labels	
149 - 150	2	N	Code Value	1st Data Code
151 - 170	20	A	Code Label	
171 - 172	2	N	Code Value	2nd Data Code
173 - 192	20	A	Code Label	
.	.	.	.	
.	.	.	.	
1381 - 1382	2	N	Code Value	57th Data Code
1383 - 1402	20	A	Code Label	

Field Sequence Number Fields are numbered sequentially to represent the order in which they appear on the raw data record.

Field Name A short name of up to eight characters that uniquely identifies the field.

Start Column The start location of the field on the data record.

End Column The end location of the field on the data record.

Field Width The number of column positions used by the field.

Decimal Places The number of digits to the right of the decimal point in the field. The raw data contain implicit decimal points.

Field Types The files include two field types:

Type 1 Discrete data with a fixed number of responses. Type 1 fields may include raw item responses or imputed categorical variables.

Type 2 Continuous numerical data without fixed ranges.

Minimum Valid Response The minimum value of valid responses for an item, excluding "I don't know" responses.

Maximum Valid Response The maximum value of valid responses for an item, excluding "I don't know" responses.

Minimum Correct Response For multiple-choice and short constructed-response cognitive items, the minimum or only correct response value.

Maximum Correct Response For short constructed-response items with more than one correct response, the maximum correct response value. For example, if possible responses for a professionally scored item ranged from 0 to 5, and 3 to 5 were considered acceptable responses, the minimum correct response would be 3 and the maximum correct response would be 5. For short constructed-response items with only one correct response and for multiple-choice cognitive items, the value in this field is the same as the Minimum Correct Response.

Illegible or Illiterate Response Code For constructed-response cognitive items, the code assigned to illegible, illiterate, or otherwise unintelligible responses, and to crossed out or erased responses.

Nonrateable Response Code For constructed-response cognitive items, the code assigned to written "I don't know" responses and off-task responses that do not address the given task.

I Don't Know Response Code For multiple-choice items, the code assigned to "I don't know" responses when that response option was given.

Omit Code The value in this field is the code assigned to nonresponses for the following types of items:

- All noncognitive items (background, attitude, and questionnaire)
- Cognitive items that are followed by valid responses to other items in the same block

Not Reached Code The value in this field is the code assigned to nonresponses to cognitive items after the last valid response in a block.

Multiple Response Code The value in this field is the code assigned to multiple-choice items for which the respondent indicated more than one response.

Name/ Description A 50-character description of the item or variable.

Alternate NAEP ID For some blocks, the identification code printed with the item text in the assessment booklet. If the item was used prior to 1983, the identification code previously assigned to the item.

Scaling Category A nonzero code in this column denotes the usage of the item described by this record on one of the two purpose-for-reading scales derived for the 1994 reading assessment. Table 9-5 lists the code values, their corresponding scales, and the name of the related scale variable(s).

Number of Response Categories A value of one (1) in this column denotes an item that was scaled by a three parameter logistic IRT model. A value greater than one signifies an item scaled by a polytomous response model and indicates the number of category parameters that were estimated for the item.

Number of Response Categories (continued)

IRT Parameters If the previous field has a value of one (1), this field contains the three IRT parameters for a dichotomous response model, "a" (slope), "b" (threshold), and "c" (asymptote). If the previous field value is greater than one, this field contains the polytomous response model parameters: "a" (slope), "b" (location), and "d" (category). The value in the previous field denotes the number of "d" parameters. Each parameter is represented to a precision of five decimal places with an explicit decimal point.

Number of Data Codes and Labels The number of valid data codes. For item responses, these include the special response codes for illegible, nonrateable, "I don't know," omitted, not reached, and multiple responses.

Data Codes and Labels For each discrete variable, a two-position field that shows the data code and a 20-position text field that provides a brief description of the code. There can be up to 57 codes; if there are fewer than 57, the remaining fields are blank.

Table 9-5
Scaling Categories and Codes

Column	Subject	Code	Scale	Scale Variables
105	Reading	1	Reading for Literary Experience (LIT)	RRPS1x
		2	Reading to Gain Information (INF)	RRPS2x

9.3 Printed Documentation

Each state's data files are accompanied by a book containing the layouts and codebooks for each data file. These documents are grouped by layout and codebook pair in the same order as the data files.

9.3.1 File Layouts

Each file layout includes the following information for each data field:

Seq. No. Sequence number. Fields are numbered sequentially to represent the order in which they appear on the data record.

Field Name A short name of up to eight characters that identifies the field. This name is used consistently across all documentation, SAS and SPSS control files, and catalog files to identify each field uniquely within a data file. In general, nonresponse data field names are abbreviations of the field descriptions. Field names associated with response data are formatted as follows:

Position 1 identifies the nature or source of the response data:

- B= Common background item within common background block
- C= School questionnaire item
- R= Reading cognitive or background item
- S= Subject-related background or attitude item (noncognitive reading items from the 1984 and 1986 assessments)
- T= Teacher questionnaire item
- X= Excluded student questionnaire item

Positions 2-5 identify an exercise (student files) or question (school, teacher, and excluded student files). Reading background items are identified by "R" in position 1 and "8" in position 2.

Positions 6-7 identify a part within an exercise (student files) or a part within a question (school, teacher, excluded student files).

Position 8 identifies the block containing an item (student files only) to avoid duplicated naming of items that occur in more than one block. The numeric block designation (2 through 18) has been replaced by an alphabetic one (B through R). This position is blank for questionnaire items and all other variables.

Col. Pos. Column position. The start location of the field on the data record.

Field Width The number of column positions used by the field.

Decimal Places The number of digits to the right of the decimal point in the field. The raw data contain implicit decimal points.

Type The files include five field types:

Type C Continuous numerical data without fixed ranges.

Type D Discrete data with a fixed number of responses. Type D fields may include raw item responses or imputed (derived) categorical variables.

Type DI Discrete data with a special code for "I don't know" responses.

Type OS Constructed-response items in the student data that were professionally scored at ETS and dichotomized for scaling.

Type OE Constructed-response items in the student data that were professionally scored at ETS and scaled as is under a polytomous response model.

Range The range of values or the range of valid responses for a field.

Key Value For multiple-choice cognitive items and for those constructed-response items that were scored using a cut-point scale, the correct response(s).

Short Label A brief description of the information in the field.

9.3.2 Codebooks

The codebook contains one or more lines of information for each data field, depending on the data type. The first line of each codebook entry contains the following information:

Seq. No. Sequence number. The fields in the codebooks are numbered sequentially and are identical to the numbers used in the layout.

Field Name A short name of up to eight characters that uniquely identifies the field. For some items, the identification code printed with the item text in the assessment booklet is printed in parentheses below the Field Name. If an item was used in an assessment prior to 1984, the identification code previously assigned to the item is located in parentheses below the Field Name.

Rel. Ind. Release indicator. A "P" indicates that an item is available for unrestricted public use. Test items that are not classified as public release are identified by a release indicator of "N."

Type In conjunction with the field types defined for the layouts, the field type is designated as continuous (C), discrete (D), discrete with "I don't know" (DI), constructed-response short (OS), or constructed-response extended (OE).

Block For assessment items, indicates the block in which an item appeared for the cohort of students for which the codebook was prepared.

Item No. Indicates the order of an item within a block for the grade of students for which the codebook was prepared.

Ages Indicates the student grade group(s) to which an item was administered: 1 = Grade 4, 2 = Grade 8, 3 = Grade 12.

Name/ Description A brief description of the information in the field.

For all discrete or constructed-response data fields, the third and subsequent lines contain each valid data value, its associated label, and the unweighted frequency of that value in the data file. (For cognitive items, the correct data values are indicated by an asterisk.) The last line under each discrete variable entry contains the "TOTAL" or sum of the frequency counts as an extra check for analyses.

If an item was used in the generation of proficiency scale scores, its scoring key, scale identification, and IRT parameter values are listed to the right of the frequency data. The first column, labeled "SCORE," contains the score value assigned to each response code for use in IRT scoring. The second column contains a three-character code for the scale on which the item was calibrated (see Table 9-5). The third column contains the IRT parameter name, with its corresponding value in the fourth column.

***Note**: To maintain item security, the text describing the responses (the data value labels) has been replaced by short descriptions for items that have not been released to the public.*

CHAPTER 10

10.1 Introduction

This chapter discusses the use of the statistical software SPSS and SAS in analyzing NAEP data. Included are procedures for creating SPSS and SAS system files, merging files using SPSS and SAS, using the jackknife procedure with SPSS and SAS to estimate standard errors, and an example using NAEP data with SPSS and SAS.

10.2 SPSS and SAS Control Statement Files

All data files in the NAEP data package are accompanied by separate control files to facilitate the creation of SPSS and SAS system files. These control files include statements for variable definitions, variable labels, missing value codes, value labels, and an optional section for creating and storing scored variables. Each set of control statements also generates unweighted descriptive statistics of the reporting variables for the related data file and a listing of the contents of the saved system file.

Users who are performing analyses using data residing on magnetic tape should be aware that the system file generation programs cannot run if both the control statement file and its corresponding data file reside on the same tape. Both SPSS and SAS will try to read a data file before they have completed processing the control statement file, which is physically impossible if both files are on the same tape. The user is advised to copy the control files to disk, as they require less storage space and the user can then edit the control statements before generating the system files.

The common features of both types of control files, as well as general guidelines, are provided below.

Variable Description — The field names are listed in the order in which they appear on the file, along with their column position and input format. If a field is numeric with no decimal places, no format is provided. Otherwise, the format is indicated by a number for the number of decimal places, or by '$' or '(A)' for a nonnumeric field.

Variable Labels — A 40- or 50-character text descriptions for each field.

Missing Values — All blank fields in the data are automatically set to the system missing value by each package. Some items had special codes assigned to "I don't know," omitted, not reached, or multiple responses (Table 9-3 in Chapter 9 provides details about these codes). Optional sections of each control statement file allow the user to instruct SPSS or SAS to treat these values as missing also.

Value Labels — All numeric fields with discrete (or categorical) values are provided with 20-character text descriptors for each value within the variable's range. The value labels, or formats, for the SAS control statements have been pooled across all samples into a file for one-time processing and loading into a SAS format library. A listing that links the field names to the SAS format names is provided with the codebooks.

Scoring

For each item with one or more correct responses, control statements are provided for creating a scored variable, its label, and its value labels. The scoring of each item is performed according to the following scheme: missing values are copied as is; correct response values are recoded to 1; all other values, including no response and "I don't know," are recoded as 0. The scoring of the omit, not reached, and "I don't know"/nonrateable values are coded separately from other incorrect responses to allow the user to edit these control statements and substitute alternate values.

Each scorable item is replaced by its scored value, along with its new value labels and missing value declarations. The entire block of scoring control statements is performed conditionally by default and will not be saved on the output system file. To save the scored variables permanently, the user must edit the control statement file and make changes to a few specified statements. It is not possible under this scheme to save both the raw and scored responses to the same item.

10.3 Creating SPSS System Files

Each SPSS control statement file is linked to its corresponding data file through the file name. To obtain the control statement file name, replace the suffix DAT in the data file name with SPS. For example, file TSR1STUD.SPS is the control statement file for data file TSR1STUD.DAT.

All SPSS control statement files have been generated according to the structure in Table 10-1.

A set of "MISSING VALUES" declarations are provided to allow the designation of "I don't know"/nonrateable, omitted, not reached, or multiple responses as user-defined missing values. These statements have been commented out to allow the user to decide which, if any, of the values are to be treated as missing values. Deleting the asterisk preceding a "MISSING VALUES" statement activates that user-defined missing value for the listed variables. If the user designates more than one code as a missing value, and a variable is referenced by more than one of those statements, SPSS will use only the code corresponding to the last occurrence of that variable.

The missing value transformations are followed by a series of RECODE scoring statements to create scored variables from cognitive item variables (see Section 10.2). The TEMPORARY command instructs SPSS to perform the subsequent scoring statements on a temporary basis and delete the new variables after the next procedure encountered (FREQUENCIES). Thus, the scored variables will NOT be saved on the system file unless the TEMPORARY command is commented or edited out.

All control statement files assume that the file handle (or DDNAME) for the input data file is RAWDATA, and the file handle for the output system file is SYSFILE.

The control statements were coded according to the command and procedure descriptions in the *SPSS® Base System Syntax Reference Guide, Release 6.0* (SPSS, Inc., 1993). They were tested under SPSS® for Windows™, Release 6.1.1.

10.4 Creating SAS System Files

Each SAS control statement file is linked to its corresponding data file through the file name. To obtain the control statement file name, replace the suffix DAT in the data file name with SAS. For example, file TSR1STUD.SAS is the control statement file for data file TSR1STUD.DAT.

All SAS control statement files have been generated according to the structure in Table 10-2.

Table 10-1
SPSS Control Statement Synopsis

```
TITLE
  label for sysout of file generation run
FILE LABEL
  label to be stored with file
DATA LIST FILE=RAWDATA
  variable names, locations, and formats
VARIABLE LABELS
  40-character label for each variable
*MISSING VALUES
  list of variables to have user-missing values assigned
VALUE LABELS
  variable names, values, and value labels
DOCUMENT
  text description of data to be saved in file
TEMPORARY.          ** delete this statement to save scored variables **
RECODE
  oldvar (SYSMIS=SYSMIS)  (0=9)  (keyval[s]=1)
  (nrval=0)  (omval=0) (idkval=0)  (ELSE=0)
.
.
.

MISSING VALUES
  for recodes of multiple responses
VALUE LABELS
  1=Correct    0=Incorrect
FREQUENCIES
  reporting variables
SAVE   OUTFILE=SYSFILE/COMPRESSED
DISPLAY LABELS.
```

Table 10-2
SAS Control Statement Synopsis

```
TITLE
DATA SYSFILE.xxx;
INFILE RAWDATA;
INPUT
  variable names, positions, and formats
LABEL
  40-character variable labels

ARRAY DKn (I)              list of variables with "I Don't
*DO OVER DKn;              Know" or unrateable codes to be recoded
*  IF DKn=7   THEN DKn=.;  for missing
*  END;
ARRAY OMn (I)              list of variables with omit codes to be
*DO OVER OMn;              recoded for missing
*  IF OMn=8   THEN OMn=.;
*  END;
ARRAY NRn (I)              list of variables with not-reached codes
*DO OVER NRn;              to be recoded for missing
*  IF NRn=9   THEN NRn=.;
*  END;
ARRAY MRn (I)              list of variables with multiple response
*DO OVER MRn;              codes to be recoded for missing
*  IF MRn=0   THEN MRn=.;
*  END;
LENGTH DEFAULT=2
  other variables with appropriate lengths;
%MACRO RECODE;
  SAS macro to perform scoring for each variable
%MEND RECODE;
%MACRO SCORE;
%RECODE (oldvar,newvar,idkval,omval,nrval,mrval,key1val,[key2val])
  .
  .
%MEND;
*%SCORE;                   ** delete asterisk to save scored variables **
RUN;
PROC FORMAT LIBRARY=SASLIB
   VALUE
          formats for the reporting variables
PROC FREQ;
TABLES
  reporting variables
PROC CONTENTS POSITION;
```

They use the SAS Macro Language facility to reduce the number of source statements generated and to provide consistent performance of repetitive functions. Therefore, the user must ensure that the MACRO option is invoked when processing any of the control statement files.

The DO OVER through END statements following each ARRAY statement set up the conversion of the "I don't know"/nonrateable, omitted, not reached, and multiple response codes to the system missing value. However, once this conversion is executed and saved on the system file, these recoded values will be indistinguishable from actual missing values on the original data file. For this reason, these statements have been commented out to allow the user to decide which, if any, of the values are to be recoded. To activate the recoding, delete the asterisks preceding the appropriate DO OVER, IF THEN, and END statements.

The missing value transformations are followed by a series of SAS macro definitions for scoring the cognitive items. The RECODE macro is used by the SCORE macro to transform the responses to each item into score values. The RECODE macro may be edited by the user to transform the special codes for each item consistently into other values.

At the end of the control statements, the SCORE macro is commented out. To save the scored variables on the system file, the user should uncomment the %SCORE statement.

A separate file of SAS control statements contains the SAS formats to be used by all variables. This file, named FORMAT1.SAS, may be executed before all other SAS control statement files, and does not require a raw data file for input. The format specifications will be saved in a library designated to the system as SASLIB. Each codebook contains a list of all discrete variables and the format values to be used in any SAS analysis.

The control statements were coded according to the command and procedure descriptions in the *SAS Language: Reference, Version 6, First Edition* (SAS Institute, Inc., 1990). They were tested under The SAS System for Windows™, Release 6.10.

10.5 Merging Files Under SPSS or SAS

The NAEP data files are structured to facilitate matching among the four instruments (student, excluded student, teacher, and school). The teacher questionnaire data have already been linked with the records of appropriate students on the state and national comparison samples. For the purposes of analysis and reporting, only two types of linkages are valid:

1) school with student and teacher (state and national)

2) school with excluded student (state only)

The principal linkage between files is accomplished through the scrambled primary sampling unit/school code field, which is named SCRPSU on the student data files, SSCRPSU on the school data files, and XSCRPSU on the excluded student data files. All files are sorted by this field to permit direct match-merging without the need to re-sort.

When a hierarchical file match is performed, both SPSS and SAS build a rectangular file at the level of the lowest file in the match. Each record from the higher order file is repeated across the corresponding records of the lower order file. For example, in matching school with student data, the information from one school record is repeated across all student records belonging to that school. Clearly, the number of variables from the higher order file will have a greater impact on the size of the resulting merged file.

The examples shown in Tables 10-3 and 10-4 will perform direct matches according to the linkages listed above. The KEEP statements are not necessary to the performance of the merge, but when they are applied to only those variables required for analysis, they will make more efficient use of computer resources. These examples also assume that no transformations are to be performed on the input files. If transformations are desired for analysis, the most efficient course to follow would be to transform the variables from the higher order file first, perform the match procedure, and then transform the variables from the lower order file.

Table 10-3
Matching School and Student Files

SPSS

```
MATCH FILES
  TABLE=SCHOOL/
        RENAME=(SSCRPSU=SCRPSU)/
  FILE=STUDENT/
        KEEP=SCRPSU, other school & student variables/
  BY=SCRPSU.
```

SAS

```
DATA MATCH1;
  MERGE SCHOOL(RENAME=(SSCRPSU=SCRPSU)
               KEEP=SCRPSU other school variables)
        STUDENT(KEEP=SCRPSU other student variables);
  BY SCRPSU;
```

Table 10-4
Matching School and Excluded Student Files, State Sample

SPSS

```
MATCH FILES
  TABLE=SCHOOL/
        RENAME=(SSCRPSU=SCRPSU)/
  FILE=EXCLUDE/
        RENAME=(XSCRPSU=SCRPSU)/
        KEEP=SCRPSU, other school and excluded student variables/
  BY=SCRPSU.
```

SAS

```
DATA MATCH3;
  MERGE SCHOOL (RENAME=(SSCRPSU=SCRPSU)
               KEEP=SCRPSU other school variables)
        EXCLUDE(RENAME=(XSCRPSU=SCRPSU)
               KEEP=SCRPSU other excluded student variables);
  BY SCRPSU;
```

10.6 Computing the Estimated Variance of a Mean (Jackknifing) Using SPSS or SAS

This section presents the two multiweight methods for computing the estimated variance of a mean in SPSS and SAS program code form (see Section 8.3 in Chapter 8 for a discussion of the jackknife procedure). The first method may be used for any variable except the plausible values. The second method, which should be used for the plausible values, employs a correction for the variance in estimating the values (correction for imputation).

For each variable to be jackknifed, generate two vectors of weighted sums and products. Sum these vectors across the entire file using the AGGREGATE (SPSS) or SUMMARY (SAS) procedures. From the weighted sums compute the weighted means and then compute the estimated variance and standard error.

One advantage to this approach is that it will accomplish the computation in one pass of the data. Another advantage, afforded by the AGGREGATE (SPSS) and SUMMARY (SAS) procedures, is the facility to compute subgroup statistics by using the BREAK keyword (SPSS) or CLASS option (SAS) with the variable(s) defining the subgroups. All computations performed subsequent to the aggregation procedure are performed on each record of the collapsed file, corresponding to one of the subgroups. In the examples in Tables 10-5, 10-6, 10-7, and 10-8, the variable DSEX (gender) is used as a break control variable, and the derived statistics are printed for each gender code.

In Tables 10-5 and 10-6, the variable X may be any variable or transformation of variables except plausible values. In Tables 10-7 and 10-8, the vector or array named VALUE refers to a set of plausible values.

Table 10-5
Standard Error Computation: Multiweight Method Using SPSS

```
GET FILE=SYSFILE/               (System file for sample)
  KEEP=DSEX,ORIGWT,SRWT01 TO SRWT62,X.
VECTOR WT=SRWT01 TO SRWT62.
VECTOR WX(62).
SELECT IF (NOT SYSMIS(X)).
COMPUTE WTX=ORIGWT*X.
LOOP #I=1 TO 62.
   COMPUTE WX(#I) = WT(#I)*X.
END LOOP.
AGGREGATE  OUTFILE=*/BREAK=DSEX/UWN=N(ORIGWT)/
  SWT,SW1 TO SW62 = SUM(ORIGWT,SRWT01 TO SRWT62)/
  SWX,SX1 TO SX62 = SUM(WTX,WX1 TO WX62)/.
VECTOR SW = SW1 TO SW62.
VECTOR SX = SX1 TO SX62.
COMPUTE XBAR = SWX/SWT.
COMPUTE XVAR = 0.
LOOP #I=1 TO 62.
   COMPUTE #DIFF = SX(#I)/SW(#I) - XBAR.
   COMPUTE XVAR = XVAR + #DIFF * #DIFF.
END LOOP.
COMPUTE XSE = SQRT(XVAR).
PRINT FORMATS XVAR,XSE (F8.4).
LIST VARIABLES=DSEX,UWN,SWT,XVAR,XSE.
FINISH.
```

Table 10-6

Standard Error Computation: Multiweight Method Using SAS

```
DATA A;
  SET SYSFILE.;                (System file for sample)
  ARRAY WT SRWT01-SRWT62;
  ARRAY WX WX1-WX62;
  IF (X NE .);
  WTX = ORIGWT*X;
  DO OVER WT;
        WX = WT*X;
  END;
PROC SUMMARY;
  CLASS DSEX;
  VAR ORIGWT SRWT01-SRWT62 WTX WX1-WX62;
  OUTPUT OUT=B    N(ORIGWT)=UWN
        SUM(ORIGWT WTX SRWT01-SRWT62 WX1-WX62)=
            SWT SWX SW1-SW62 SX1-SX62;
DATA C;
  SET B;
  ARRAY SW SW1-SW62;
  ARRAY SX SX1-SX62;
  XBAR = SWX/SWT;
  XVAR = 0;
  DO OVER SW;
        DIFF = (SX/SW)-XBAR;
        XVAR = XVAR+DIFF*DIFF;
  END;
  XSE = SQRT(XVAR);
PROC PRINT;
  VAR DSEX UWN SWT XVAR XSE;
```

Table 10-7

Standard Error Computation: Multiweight Method Using SPSS with Correction for Imputation

```
GET FILE=SYSFILE/               (System file for sample)
  KEEP=DSEX,ORIGWT,SRWT01 TO SRWT62,X.
VECTOR VALUE=RRPCM1 TO RRPCM5.
VECTOR WT=SRWT01 TO SRWT62.
VECTOR WX(62).
VECTOR WS(5).
SELECT IF (NOT SYSMIS(RRPCM1)).
COMPUTE WTX=ORIGWT*RRPCM1.
LOOP #I=1 TO 62.
   COMPUTE WX(#I) = WT(#I)*RRPCM1.
END LOOP.
LOOP #I=1 TO 5.
  COMPUTE WS(#I) = VALUE(#I)*ORIGWT.
END LOOP.
AGGREGATE  OUTFILE=*/BREAK=DSEX/UWN=N(ORIGWT)/
  SWT,SW1 TO SW62 = SUM(ORIGWT,SRWT01 TO SRWT62)/
  SWX,SX1 TO SX62 = SUM(WTX,WX1 TO WX62)/
  SS1 TO SS5 = SUM(WS1 TO WS5)/.
VECTOR SW = SW1 TO SW62.
VECTOR SX = SX1 TO SX62.
VECTOR SS = SS1 TO SS5.
COMPUTE XBAR = SWX/SWT.
COMPUTE XVAR = 0.
LOOP #I=1 TO 62.
   COMPUTE #DIFF = SX(#I)/SW(#I) - XBAR.
   COMPUTE XVAR = XVAR + #DIFF * #DIFF.
END LOOP.
LOOP #I=1 TO 5.
  COMPUTE SS(#I) = SS(#I)/SWT.
END LOOP.
  COMPUTE SBAR = MEAN(SS1 TO SS5).
COMPUTE SVAR = VARIANCE(SS1 TO SS5).
COMPUTE XSE = SQRT(XVAR+(6/5)*SVAR).
PRINT FORMATS SBAR,XVAR,SVAR,XSE (F8.4).
LIST VARIABLES=DSEX,UWN,SWT,SBAR,XVAR,SVAR,XSE.
FINISH.
```

Table 10-8
Standard Error Computation: Multiweight Method Using SAS with Correction for Imputation

```
DATA A;
  SET SYSFILE.;               (System file for sample)
  ARRAY WT SRWT01-SRWT62;
  ARRAY WX WX1-WX62;
  ARRAY VALUE RRPCM1-RRPCM5;
  ARRAY WS WS1-WS5;
  IF (RRPCM1 NE .);
  WTX = ORIGWT*RRPCM1;
  DO OVER WT;
        WX = WT*RRPCM1;
  END;
  DO OVER WS;
        WS = VALUE*ORIGWT;
  END;
PROC SUMMARY;
  CLASS DSEX;
  VAR ORIGWT SRWT01-SRWT62 WTX WX1-WX62 WS1-WS5;
  OUTPUT OUT=B    N(ORIGWT)=UWN
        SUM(ORIGWT WTX SRWT01-SRWT62 WX1-WX62 WS1-WS5)=
            SWT SWX SW1-SW62 SX1-SX62 SS1-SS5;
DATA C;
  SET B;
  ARRAY SW SW1-SW62;
  ARRAY SX SX1-SX62;
  ARRAY SS SS1-SS5;
  XBAR = SWX/SWT;
  XVAR = 0;
  DO OVER SW;
        DIFF = (SX/SW)-XBAR;
        XVAR = XVAR+DIFF*DIFF;
  END;
  DO OVER SS;
        SS = SS/SWT;
  END;
        SBAR = MEAN(SS1,SS2,SS3,SS4,SS5);
  SVAR = VAR(SS1,SS2,SS3,SS4,SS5);
  XSE = SQRT(XVAR+(6/5)*SVAR);
PROC PRINT;
  VAR DSEX UWN SWT SBAR XVAR SVAR XSE;
```

10.7 An Analysis Example Using NAEP Data with SPSS and SAS

In Chapter 1, we explained how to perform an analysis of NAEP data using any statistical or procedural language, and presented an example of how to produce a simple descriptive analysis table that did not include standard error estimates.

This section explains how to use SPSS and SAS to perform the same analysis, this time including standard error estimates that account for NAEP sampling design and measurement error components. Such an accounting is required for statistical comparison of NAEP data. Because the NAEP sample is not a simple random sample, ordinary formulas for estimating the standard error of sample statistics will produce values that are too small.

Before attempting any analysis of NAEP data, users should understand the special characteristics of the NAEP sampling design, described in Chapters 2 and 4. Alternate methods for computing standard errors and recommended formulas for obtaining degrees of freedom are given in Chapter 8.

The analysis in our example produced the following estimates, with standard errors, of the reported amount of television watched each day by fourth-grade public school girls in the national comparison sample and the corresponding mean reading proficiency scores. The output from SPSS is given in Table 10-9, while output from SAS is shown in Table 10-10. Similar tables for each state are included at the beginning of each state's codebook.

Table 10-9

SPSS Analysis Example Using Jackknife Standard Error Estimates

1994 National Comparison Sample
Reading Results for 4th Grade Public-School Girls
by Amount of Television Viewing

HOW MUCH TELEVISION DO YOU USUALLY WATCH	N	WTD N	PCT	SE(PCT)	MEAN	SE(MEAN)
NONE	45	24226.55	1.579	.28753	220.322	9.77326
1 HOUR OR LESS	572	300946.57	19.612	.97512	220.677	2.52593
2 HOURS	620	343473.39	22.383	.87664	223.968	1.69413
3 HOURS	491	265853.73	17.325	.92766	225.063	1.79654
4 HOURS	367	199259.39	12.985	.86173	225.758	2.58381
5 HOURS	249	126470.41	8.242	.67875	213.613	3.14221
6 HOURS OR MORE	611	274272.98	17.874	.93823	197.146	2.22712

Table 10-10
SAS Analysis Example Using Jackknife Standard Error Estimates

1994 National Comparison Sample
Reading Results for 4th Grade Public-School Girls
by Amount of Television Viewing

OBS	HOW MUCH TELEVISION DO YOU USUALLY WATCH	N	WTD N	PCT	SE(PCT)	MEAN	SE(MEAN)
1	TOTAL	2955	1534503.02	100.000	0.00000	218.040	1.19310
2	NONE	45	24226.55	1.579	0.28753	220.322	9.77326
3	1 HOUR OR LESS	572	300946.57	19.612	0.97512	220.677	2.52593
4	2 HOURS	620	343473.39	22.383	0.87664	223.968	1.69413
5	3 HOURS	491	265853.73	17.325	0.92766	225.063	1.79654
6	4 HOURS	367	199259.39	12.985	0.86173	225.758	2.58381
7	5 HOURS	249	126470.41	8.242	0.67875	213.613	3.14221
8	6 HOURS OR MORE	611	274272.98	17.874	0.93823	197.146	2.22712

Most analyses of NAEP data can be performed in four basic steps:

- Identify and access the appropriate data file
- Identify and extract the relevant variables
- Select the proper subset of students
- Compute and print the results

The method you choose to perform these steps may vary with the complexity of the analysis or with the statistical or procedural language you are using.

To begin the example analysis, you need to identify

- the national file that contains response data for fourth-grade students and
- the relevant variables in the file.

NAEP files are described in Chapter 9 and listed in Table 9-1; the correct file for our example is NCR1STUD.DAT. Next, find the data set record layout for NCR1STUD.DAT in the accompanying document entitled *Layouts and Codebooks*. Here you will find the names and file locations of the variables needed to produce this table (response counts for each variable are found in the corresponding codebook). To produce the table, we need four variables for the basic data and weights, 62 replicate weight variables to produce the standard errors, and five plausible values (Table 10-11).

For analyses that are relatively simple (requiring the use of just a few variables), you can manually enter the variable labels and locations into your computer program. This example can be performed more efficiently through the use of the SPSS or SAS control statement files.

As an aid to users, three types of files have been included:

- machine-readable catalog files
- SPSS control statement files
- SAS control statement files

The SPSS and SAS control statement files are provided to facilitate the creation of SPSS and SAS system files. There is an SPSS and a SAS control file for the data file(s) for each sample (see Table 9-1). Part of each control file contains the field name, location, and format for each variable on the corresponding data file. More about control statement files can be found in Sections 10.2 through 10.4.

Table 10-11
NAEP Variables Used to Produce the Analysis

Seq. No.	Field Name	Column Position	Field Width	Decimal Places	Type	Range	Short Label
28	SCHTYPE	68	1	–	D	1-5	School type
36	DSEX	94	1	–	D	1-2	Gender
50	ORIGWT	175	7	2	C	–	Student weight (unadjusted)
51	SRWT01	182	7	2	C	–	Student replicate weight 01
⋮	⋮	⋮	⋮	⋮	⋮		⋮
112	SRWT62	609	7	2	C	–	Student replicate weight 62
213	RRPCM1	896	5	2	C	–	Plausible NAEP reading value #1 (Composite)
⋮	⋮	⋮	⋮	⋮	⋮		⋮
217	RRPCM5	916	5	2	C	–	Plausible NAEP reading value #5 (Composite)
229	B001801A	932	1	–	D	1-7	How much television do you usually watch each day?

Any statistical computing language or package can be used to access the data file, extract the relevant variables, select the proper subset of students, and compute the values shown in the table. Using SPSS or SAS, the following procedure will complete the analysis example.

1) Select the file containing the appropriate grade 4 student data. This is the national sample student data file described in Table 9-1; its file name is NCR1STUD.DAT. Identify the relevant variables from the data set record layout: SCHTYPE, DSEX, ORIGWT, SRWT01-SRWT62, B001801A, and RRPCM1-RRPCM5.

2) Using the raw data file, select the appropriate subset of students for the table. This selection restricts the analysis to public-school (SCHTYPE=1) girls (DSEX=2) who have valid RRPCM1 (reading proficiency) and B001801A (television viewing) values.

3) Compute weighted products and sums corresponding to the 62 student replicate weights and the five estimates of student reading proficiency.

4) Compute overall weighted sums for use in the computation of percentages and jackknife standard errors.

5) Compute weighted sums for each level of television viewing (B001801A).

6) Merge the weighted sums from steps 4 and 5 and compute percentages, variances, and jackknife standard errors (with sampling and measurement error components).

7) Print the final result in a formatted table.

The SPSS code for performing steps 1 to 7 is shown in Table 10-12; the SAS code for performing steps 1 to 7 is shown in Table 10-13.

Table 10-12

SPSS Code for Steps 2 through 7 to Produce Example Analysis

```
TITLE      "1994 National Comparison Sample:  Reading Results for".
SUBTITLE   "4th Grade Public-School Girls by Amount of TV Viewing".
FILE HANDLE  NCR1STUD  /NAME='G:\DATA\NCR1STUD.DAT' /LRECL=1524.
* ----------------------------- STEP 1 ----------------------------- .
DATA LIST FILE=NCR1STUD/
  DSEX        94         SCHTYPE     68
  ORIGWT     175-181 (2) SRWT01     182-188 (2) SRWT02      189-195 (2)
  SRWT03     196-202 (2) SRWT04     203-209 (2) SRWT05      210-216 (2)
  SRWT06     217-223 (2) SRWT07     224-230 (2) SRWT08      231-237 (2)
  SRWT09     238-244 (2) SRWT10     245-251 (2) SRWT11      252-258 (2)
  SRWT12     259-265 (2) SRWT13     266-272 (2) SRWT14      273-279 (2)
  SRWT15     280-286 (2) SRWT16     287-293 (2) SRWT17      294-300 (2)
  SRWT18     301-307 (2) SRWT19     308-314 (2) SRWT20      315-321 (2)
  SRWT21     322-328 (2) SRWT22     329-335 (2) SRWT23      336-342 (2)
  SRWT24     343-349 (2) SRWT25     350-356 (2) SRWT26      357-363 (2)
  SRWT27     364-370 (2) SRWT28     371-377 (2) SRWT29      378-384 (2)
  SRWT30     385-391 (2) SRWT31     392-398 (2) SRWT32      399-405 (2)
  SRWT33     406-412 (2) SRWT34     413-419 (2) SRWT35      420-426 (2)
  SRWT36     427-433 (2) SRWT37     434-440 (2) SRWT38      441-447 (2)
  SRWT39     448-454 (2) SRWT40     455-461 (2) SRWT41      462-468 (2)
  SRWT42     469-475 (2) SRWT43     476-482 (2) SRWT44      483-489 (2)
  SRWT45     490-496 (2) SRWT46     497-503 (2) SRWT47      504-510 (2)
  SRWT48     511-517 (2) SRWT49     518-524 (2) SRWT50      525-531 (2)
  SRWT51     532-538 (2) SRWT52     539-545 (2) SRWT53      546-552 (2)
  SRWT54     553-559 (2) SRWT55     560-566 (2) SRWT56      567-573 (2)
  SRWT57     574-580 (2) SRWT58     581-587 (2) SRWT59      588-594 (2)
  SRWT60     595-601 (2) SRWT61     602-608 (2) SRWT62      609-615 (2)
  B001801A   932         RRPCM1     896-900 (2) RRPCM2      901-905 (2)
  RRPCM3     906-910 (2) RRPCM4     911-915 (2) RRPCM5      916-920 (2).
VECTOR VALUE=RRPCM1 TO RRPCM5.
VECTOR    WT=SRWT01 TO SRWT62.
VECTOR    WX(62).
VECTOR    WS(62).
* ----------------------------- STEP 2 ----------------------------- .
SELECT IF (NOT SYSMIS(RRPCM1)).
SELECT IF DSEX = 2.
SELECT IF SCHTYPE = 1.
SELECT IF RANGE(B001801A,1,7).
* ----------------------------- STEP 3 ----------------------------- .
COMPUTE WTX = ORIGWT*RRPCM1.
LOOP #I = 1 TO 62.
.   COMPUTE WX(#I) = WT(#I)*RRPCM1.
END LOOP.
LOOP #I = 1 TO 5.
.   COMPUTE WS(#I) = VALUE(#I)*ORIGWT.
END LOOP.
VARIABLE LABELS
  DSEX       'GENDER                                  '
  ORIGWT     'OVERALL STUDENT FULL-SAMPLE WEIGHT      '
  SRWT01     'STUDENT REPLICATE WEIGHT 01             '
  B001801A   'HOW MUCH TELEVISION DO YOU USUALLY WATCH'
  RRPCM1     'PLAUSIBLE NAEP READING VALUE #1 (COMP.) '.
```

(continued)

Table 10-12 (continued)

SPSS Code for Steps 2 through 7 to Produce Example Analysis

```
VALUE LABELS
  B001801A                                    1 'NONE                '
                  2 '1 HOUR OR LESS      '    3 '2 HOURS             '
                  4 '3 HOURS             '    5 '4 HOURS             '
                  6 '5 HOURS             '    7 '6 HOURS OR MORE     '.
* ------------------------------ STEP 4 ------------------------------ .
AGGREGATE   OUTFILE=TEMP1/ BREAK=DSEX/
      TOTSW,TOTSW01 TO TOTSW62 = SUM(ORIGWT,SRWT01 TO SRWT62).
* ------------------------------ STEP 5 ------------------------------ .
AGGREGATE   OUTFILE=*/ BREAK=DSEX B001801A/ UWN = N(ORIGWT)/
      SWT,SW1 TO SW62 = SUM(ORIGWT,SRWT01 TO SRWT62)/
      SWX,SX1 TO SX62 = SUM(WTX,WX1 TO WX62)/
      SS1 TO SS5 = SUM(WS1 TO WS5).
* ------------------------------ STEP 6 ------------------------------ .
MATCH FILES  FILE=* / TABLE=TEMP1/ BY DSEX.
VECTOR    SW = SW1 TO SW62.
VECTOR    SX = SX1 TO SX62.
VECTOR   TSW = TOTSW01 TO TOTSW62.
VECTOR    SS = SS1 TO SS5.
COMPUTE  XVAR=0.
COMPUTE  XBAR=SWX/SWT.
COMPUTE  PVAR=0.
COMPUTE  PBAR=100*(SWT/TOTSW).
LOOP #I = 1 TO 62.
.   COMPUTE #XDIFF = (SX(#I)/SW(#I)) - XBAR.
.   COMPUTE XVAR = XVAR+#XDIFF*#XDIFF.
.   COMPUTE #PDIFF = 100*(SW(#I)/TSW(#I)) - PBAR.
.   COMPUTE PVAR = PVAR+#PDIFF*#PDIFF.
END LOOP.
LOOP #I = 1 TO 5.
.   COMPUTE SS(#I) = SS(#I)/SWT.
END LOOP.
COMPUTE SBAR=MEAN(SS1 TO SS5).
COMPUTE SVAR=VARIANCE(SS1 TO SS5).
COMPUTE SSE = SQRT(XVAR+(6/5)*SVAR).
COMPUTE PSE = SQRT(PVAR).
PRINT FORMATS  SWT (F10.2)  PBAR SBAR (F10.3)  PSE SSE (F11.5).
* ------------------------------ STEP 7 ------------------------------ .
REPORT
  / FORMAT = LIST AUTOMATIC ALIGN(CENTER) MARGINS(1,121)
  / TITLE = CENTER
       '1994 National Comparison Sample'
       'Reading Results for 4th Grade Public-School Girls'
       'by Amount of Television Viewing'
  / VARIABLES = B001801A (LABEL)  UWN 'N'  SWT 'WTD N'
      PBAR 'PCT'  PSE 'SE(PCT)'  SBAR 'MEAN'  SSE 'SE(MEAN)'.
```

Table 10-13

SAS Code for Steps 2 through 7 to Produce Example Analysis

```
TITLE1 '1994 National Comparison Sample';
TITLE2 'Reading Results for 4th Grade Public-School Girls';
TITLE3 'by Amount of Television Viewing';
/****************************** STEP 1 ******************************/
DATA    A;
INFILE  'G:\DATA\NCR1STUD.DAT'  LRECL=1524;
INPUT
  DSEX         94        SCHTYPE      68
  ORIGWT      175-181 .2 SRWT01      182-188 .2 SRWT02      189-195 .2
  SRWT03      196-202 .2 SRWT04      203-209 .2 SRWT05      210-216 .2
  SRWT06      217-223 .2 SRWT07      224-230 .2 SRWT08      231-237 .2
  SRWT09      238-244 .2 SRWT10      245-251 .2 SRWT11      252-258 .2
  SRWT12      259-265 .2 SRWT13      266-272 .2 SRWT14      273-279 .2
  SRWT15      280-286 .2 SRWT16      287-293 .2 SRWT17      294-300 .2
  SRWT18      301-307 .2 SRWT19      308-314 .2 SRWT20      315-321 .2
  SRWT21      322-328 .2 SRWT22      329-335 .2 SRWT23      336-342 .2
  SRWT24      343-349 .2 SRWT25      350-356 .2 SRWT26      357-363 .2
  SRWT27      364-370 .2 SRWT28      371-377 .2 SRWT29      378-384 .2
  SRWT30      385-391 .2 SRWT31      392-398 .2 SRWT32      399-405 .2
  SRWT33      406-412 .2 SRWT34      413-419 .2 SRWT35      420-426 .2
  SRWT36      427-433 .2 SRWT37      434-440 .2 SRWT38      441-447 .2
  SRWT39      448-454 .2 SRWT40      455-461 .2 SRWT41      462-468 .2
  SRWT42      469-475 .2 SRWT43      476-482 .2 SRWT44      483-489 .2
  SRWT45      490-496 .2 SRWT46      497-503 .2 SRWT47      504-510 .2
  SRWT48      511-517 .2 SRWT49      518-524 .2 SRWT50      525-531 .2
  SRWT51      532-538 .2 SRWT52      539-545 .2 SRWT53      546-552 .2
  SRWT54      553-559 .2 SRWT55      560-566 .2 SRWT56      567-573 .2
  SRWT57      574-580 .2 SRWT58      581-587 .2 SRWT59      588-594 .2
  SRWT60      595-601 .2 SRWT61      602-608 .2 SRWT62      609-615 .2
  B001801A    932        RRPCM1      896-900 .2 RRPCM2      901-905 .2
  RRPCM3      906-910 .2 RRPCM4      911-915 .2 RRPCM5      916-920 .2;
ARRAY WT     SRWT01-SRWT62;
ARRAY WX     WX1-WX62;
ARRAY VALUE  RRPCM1-RRPCM5;
ARRAY WS     WS1-WS5;
/****************************** STEP 2 ******************************/
IF (RRPCM1 NE .);
IF (SCHTYPE EQ 1);
IF (DSEX  EQ 2);
IF (B001801A NE .) AND
   (B001801A GT 0) AND
   (B001801A LT 8);
WTX = ORIGWT*RRPCM1;
/****************************** STEP 3 ******************************/
DO OVER WT;
     WX = WT*RRPCM1;
END;
DO OVER WS;
     WS = VALUE*ORIGWT;
END;
MDUMMY = 0;
KEEP  ORIGWT DSEX B001801A SRWT01-SRWT62 RRPCM1-RRPCM5
      WX1-WX62 WS1-WS5 WTX MDUMMY;
```

(continued)

Table 10-13 (continued)

SAS Code for Steps 2 through 7 to Produce Example Analysis

```
LABEL
  DSEX     = 'GENDER                                        '
  ORIGWT   = 'STUDENT WEIGHT (UNADJUSTED)                   '
  SRWT01   = 'STUDENT REPLICATE WEIGHT 01                   '
  B001801A = 'HOW MUCH TELEVISION DO YOU USUALLY WATCH'
  RRPCM1   = 'PLAUSIBLE NAEP READING VALUE #1 (COMP.) ';
PROC FORMAT;
  VALUE DSEX      1='MALE                '    2='FEMALE              ';
  VALUE B001801A  .='TOTAL               '    1='NONE                '
                  2='1 HOUR OR LESS      '    3='2 HOURS             '
                  4='3 HOURS             '    5='4 HOURS             '
                  6='5 HOURS             '    7='6 HOURS OR MORE     ';
/****************************** STEP 4 ******************************/
PROC SUMMARY;
    VAR MDUMMY ORIGWT SRWT01-SRWT62;
    OUTPUT OUT=B   SUM(MDUMMY)=MDUMMY
        SUM(ORIGWT SRWT01-SRWT62) = TOTSWT TOTSW1-TOTSW62;
/****************************** STEP 5 ******************************/
PROC SUMMARY DATA=A;
    CLASS B001801A;
    VAR ORIGWT SRWT01-SRWT62
        WTX WX1-WX62 WS1-WS5
        MDUMMY;
    OUTPUT OUT=C    N(ORIGWT)=UWN
        N(SRWT01-SRWT62) = NSW1-NSW62
        SUM(ORIGWT SRWT01-SRWT62 WTX WX1-WX62 WS1-WS5) =
            SWT    SW1-SW62      SWX SX1-SX62 SS1-SS5
        SUM(MDUMMY) = MDUMMY;
/****************************** STEP 6 ******************************/
DATA D;
    MERGE B C;
    BY MDUMMY;
    ARRAY SW     SW1-SW62;
    ARRAY TOTSW  TOTSW1-TOTSW62;
    ARRAY SX     SX1-SX62;
    ARRAY SS     SS1-SS5;
    P = 100.0*SWT/TOTSWT;
    XBAR = SWX/SWT;
    XVAR = 0;
    DO OVER SW;
         DIFF = (SX/SW)-XBAR;
         XVAR = XVAR+DIFF*DIFF;
    END;
    DO OVER SS;
         SS = SS/SWT;
    END;
    SBAR = MEAN(SS1,SS2,SS3,SS4,SS5);
    SVAR = VAR(SS1,SS2,SS3,SS4,SS5);
    XSE = SQRT(XVAR+(6/5)*SVAR);
    PSUM = 0;
```

(continued)

Table 10-13 (continued)

SAS Code for Steps 2 through 7 to Produce Example Analysis

```
    DO OVER SW;
         DIFF = 100.0*(SW/TOTSW)-P;
         PSUM = PSUM+DIFF*DIFF;
    END;
    SE = SQRT(PSUM);
/****************************** STEP 7 ******************************/
PROC PRINT SPLIT='*';
    FORMAT B001801A B001801A.;
    LABEL UWN = 'N'
          SWT = 'WTD  N'
          P   = 'PCT'
          SE  = 'SE(PCT)'
          SBAR= 'MEAN'
          XSE = 'SE(MEAN)';
    VAR  B001801A UWN SWT P SE SBAR XSE;
RUN;
```

APPENDIX A

NAEP HISTORY

The National Assessment of Educational Progress (NAEP) is a continuing, congressionally mandated national survey of the knowledge, skills, understandings, and attitudes of young Americans in major subject areas usually taught in school. Its primary goals are to detect and report the current status of, as well as changes in, the educational attainments of young Americans, and to report long-term trends in those attainments. The purpose of NAEP is to gather information that will aid educators, legislators, and others in improving the educational experience of youth in the United States. It is the first ongoing effort to obtain comprehensive and dependable achievement data on a national basis in a uniform, scientific manner.

Between 1964 and 1969, initial assessment planning and development activities were conducted for NAEP with support from both the Carnegie Corporation and the Ford Foundation. During this time, objectives and exercises were developed for many of the subject areas, sampling and data collection strategies were planned, and data analysis plans were formulated and outlined.

From its inception, NAEP has developed assessments through a consensus process. Educators, scholars, and laypersons design objectives for each subject area, proposing general goals they think Americans should achieve in the course of their education. After careful reviews, the objectives are given to item writers, who develop measurement instruments appropriate to the objectives.

After the items pass extensive reviews by subject matter specialists, measurement experts, and laypersons and are pretested in a sample of schools throughout the country, they are administered to a stratified multistage national probability sample. The young people sampled are selected so that assessment results may be generalized to the entire national population.

NAEP collected data for the first time in 1969. Since that time, samples have included over one million 9-, 13- and 17-year-old students and, as funding would allow, 17-year-olds who had left school and adults 26 to 35 years of age. In 1984, grade samples of students were added to the assessment. As Table A-1 illustrates, assessments have focused on traditional subject areas such as reading, writing, mathematics, science, and U.S. history and on less traditional areas such as citizenship, art, literature, music, computer competence, and career and occupational development.

Since 1971, NAEP has been solely supported by federal funds. Funding agencies have included the Office of Education, the National Center for Education, and the National Institute of Education. NAEP is currently supported by the U.S. Department of Education's Office of Educational Research and Improvement, National Center for Education Statistics.

NAEP was administered by the Education Commission of the States (ECS) through 1982. In 1983, Educational Testing Service (ETS) assumed responsibility for administration of the project, incorporating an updated sampling design and, at the same time, making a concerted effort to provide continuity with previous assessments.

Secondary-use data files were first produced in 1975, allowing outside researchers access to the NAEP database. In June 1985, ETS produced its first secondary-use data files, in a new format, for the 1984 assessment. This format, which has been used to produce all secondary-use data files since 1984, makes the NAEP data files easier to use (e.g., files have been more simply organized, documentation has been improved and made more accessible).

Table A-1

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1994

		Grades/Ages Assessed										
Assessment Year	**Subject Area(s)**	**Grade 3**	**Grade 4**	**Age 9**	**Grade 7**	**Grade 8**	**Age 13**	**Grade 11**	**Grade 12**	**Age 17**	**Age* 17OS**	**Adult**
1969-70	Science			X			X			X	X	X
1970-71	Reading			X			X			X	X	X
	Literature			X			X			X	X	X
1971-72	Music			X			X			X	X	X
	Social Studies			X			X			X	X	X
1972-73	Science			X			X			X	X	X
	Mathematics			X			X			X	X	X
1973-74	Career and Occupational Development			X			X			X	X	X
	Writing			X			X			X	X	
1974-75	Reading			X			X			X	X	
	Art			X			X			X	X	
	Index of Basic Skills									X	X	
1975-76	Citizenship/Social Studies			X			X			X	X	
	Mathematics**						X			X	X	
1976-77	Science			X			X			X		
	Basic Life Skills**									X		
	Science, Reading, Health**											X
1977-78	Mathematics			X			X			X		
	Consumer Skills**									X		
1978-79	Writing, Art, and Music			X			X			X		
1979-80	Reading/Literature			X			X			X	X	
	Art						X					
1981-82	Science**			X			X			X		
	Mathematics and Citizenship/Social Studies			X			X			X		
1984	Reading		X	X		X	X	X		X		
	Writing		X	X		X	X	X		X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)		X	X		X	X	X		X		
1985	Adult Literacy**											X

**Note*: *Age 17 students who had dropped out of school or had graduated prior to assessment.*

***Note*: *Small, special-interest assessment conducted on limited samples at specific grades or ages.*

Table A-1 (continued)
National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1994

		Grades/Ages Assessed									
Assessment Year	**Subject Area(s)**	**Grade 3**	**Grade 4**	**Age 9**	**Grade 7**	**Grade 8**	**Age 13**	**Grade 11**	**Grade 12**	**Age 17**	**Age* 17OS**	**Adult**
1986	Reading	X		X	X		X	X		X		
	Mathematics	X		X	X		X	X		X		
	Science	X		X	X		X	X		X		
	Computer Competence	X		X	X		X	X		X		
	U.S. History**							X		X		
	Literature**		X	X		X	X	X		X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Mathematics (long-term trend)		X	X		X	X	X		X		
	Science (long-term trend)											
1988	Reading		X	X		X	X		X	X		
	Writing		X	X		X	X		X	X		
	Civics		X	X		X	X		X	X		
	U.S. History		X	X		X	X		X	X		
	Document Literacy**					X	X		X	X		
	Geography**		X	X		X	X	X		X		
	Reading (long-term trend)		X	X		X	X	X		X		
	Writing (long-term trend)			X			X	X		X		
	Mathematics (long-term trend)			X			X			X		
	Science (long-term trend)											
	Civics (long-term trend)											

**Note*: Age 17 students who had dropped out of school or had graduated prior to assessment.*
***Note*: Small, special-interest assessment conducted on limited samples at specific grades or ages.*

Assessment Year	Subject Area(s)	Grades/Ages Assessed										
		Grade 3	Grade 4	Age 9	Grade 7	Grade 8	Age 13	Grade 11	Grade 12	Age 17	Age* 17OS	Adult
1990	Reading		X	X		X	X		X	X		
	Mathematics		X	X		X	X		X	X		
	Science		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
			X	X		X	X	X		X		
	Writing (long-term trend)			X			X			X		
				X			X			X		
	Mathematics (long-term trend)					X						
	Science (long-term trend)											
	Trial State Mathematics											
1992	Reading		X	X		X	X		X	X		
	Writing		X	X		X	X		X	X		
	Mathematics		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X			X		
			X	X		X	X			X		
	Writing (long-term trend)			X			X	X		X		
				X			X	X		X		
	Mathematics (long-term trend)		X			X						
			X									
	Science (long-term trend)											
	Trial State Mathematics											
	Trial State Reading											
1994	Reading		X	X		X	X		X	X		
	U.S. History		X	X		X	X		X	X		
	Geography		X	X		X	X		X	X		
	Reading (long-term trend)		X	X		X	X	X		X		
			X	X		X	X	X		X		
	Writing (long-term trend)			X			X			X		
				X			X			X		
	Mathematics (long-term trend)		X									
	Science (long-term trend)											
	Trial State Reading											

**Note*: Age 17 students who had dropped out of school or had graduated prior to assessment.*

***Note*: Small, special-interest assessment conducted on limited samples at specific grades or ages.*

APPENDIX B

IRT PARAMETERS FOR READING ITEMS

This appendix contains four tables of IRT (item response theory) parameters for the items that were used in each reading scale for the 1994 fourth-grade Trial State Assessment and the 1994 fourth-grade national assessment.

For each of the binary scored items used in scaling (i.e., multiple-choice items and short constructed-response items), the tables provide estimates of the IRT parameters (which corresponds to a_j, b_j, and c_j in equation (7.1) in Chapter 7) and their associated standard errors (S.E.) of the estimates. For each of the polytomously scored items (i.e., the extended constructed-response items), the tables also show the estimates of the d_{jv} parameters (see equation (7.3) in Chapter 7) and their associated standard errors.

The tables also show the block in which each item appears (*Block*) and the position of each item within its block (*Item*).

Note that the item parameters in this appendix are in the metrics used for the original calibration of the scales. The transformations needed to represent these parameters in terms of the metrics of the final reporting scales are given in Chapter 7.

Tables B-1 and B-2 contain the fourth-grade Trial State Assessment IRT parameters respectively for the Reading for Literary Experience and Reading to Gain Information scales. Tables B-3 and B-4 contain the corresponding IRT parameters for the 1994 national assessment.

Table B-1
IRT Parameters for the Trial State Assessment
Reading for Literary Experience Scale, Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R012001	RC	1	1.537	(0.066)	0.567	(0.019)	0.129	(0.009)		
R012002	RC	2	1.367	(0.034)	-0.186	(0.015)	0.000	(0.000)		
R012003	RC	3	1.846	(0.072)	-0.610	(0.027)	0.207	(0.017)		
R012004	RC	4	0.752	(0.024)	0.337	(0.022)	0.000	(0.000)		
R012005	RC	5	1.124	(0.049)	0.061	(0.033)	0.184	(0.016)		
R012006	RC	6	0.482	(0.013)	0.748	(0.019)	0.000	(0.000)	0.319	(0.039)
									-0.076	(0.047)
									-0.244	(0.055)
R012007	RC	7	0.620	(0.033)	-0.935	(0.101)	0.176	(0.033)		
R012008	RC	8	0.569	(0.021)	-0.588	(0.036)	0.000	(0.000)		
R012009	RC	9	1.258	(0.065)	-0.781	(0.060)	0.354	(0.027)		
R012010	RC	10	1.066	(0.031)	-0.438	(0.022)	0.000	(0.000)		
R012011	RC	11	1.718	(0.076)	-0.358	(0.032)	0.238	(0.018)		
R012101	RD	1	1.712	(0.079)	-1.060	(0.040)	0.299	(0.023)		
R012102	RD	2	0.662	(0.021)	-0.102	(0.024)	0.000	(0.000)		
R012103	RD	3	1.246	(0.048)	-0.639	(0.038)	0.188	(0.019)		
R012104	RD	4	0.659	(0.021)	-0.354	(0.027)	0.000	(0.000)		
R012105	RD	5	0.810	(0.041)	-0.084	(0.056)	0.185	(0.022)		
R012106	RD	6	0.884	(0.025)	0.054	(0.019)	0.000	(0.000)		
R012107	RD	7	1.346	(0.064)	0.228	(0.030)	0.241	(0.014)		
R012108	RD	8	0.661	(0.022)	-1.208	(0.043)	0.000	(0.000)		
R012109	RD	9	0.509	(0.020)	-1.594	(0.064)	0.000	(0.000)		
R012110	RD	10	0.826	(0.045)	-1.175	(0.101)	0.288	(0.038)		
R012111	RD	11	0.895	(0.023)	1.417	(0.020)	0.000	(0.000)	1.086	(0.020)
									-1.086	(0.052)
R012112	RD	12	0.734	(0.027)	-0.823	(0.039)	0.000	(0.000)		
R012601	RE	1	0.863	(0.031)	1.150	(0.032)	0.000	(0.000)		
R012602	RE	2	1.573	(0.083)	1.285	(0.029)	0.175	(0.007)		
R012603	RE	3	1.524	(0.066)	0.147	(0.025)	0.215	(0.013)		
R012604	RE	4	1.180	(0.038)	1.053	(0.023)	0.000	(0.000)		
R012605	RE	5	0.959	(0.086)	1.095	(0.046)	0.302	(0.015)		
R012606	RE	6	1.580	(0.086)	0.424	(0.028)	0.321	(0.013)		
R012607	RE	7	0.900	(0.024)	1.274	(0.017)	0.000	(0.000)	0.686	(0.021)
									-0.045	(0.031)
									-0.641	(0.057)
R012608	RE	8	0.556	(0.042)	-0.713	(0.164)	0.304	(0.045)		
R012609	RE	9	1.248	(0.091)	0.856	(0.034)	0.276	(0.014)		
R012610	RE	10	1.671	(0.109)	0.621	(0.029)	0.363	(0.013)		
R012611	RE	11	0.796	(0.027)	0.214	(0.023)	0.000	(0.000)		
R015801	RI	1	0.968	(0.045)	-1.534	(0.081)	0.241	(0.034)		
R015802	RI	2	0.405	(0.017)	-1.275	(0.062)	0.000	(0.000)		
R015803	RI	3	0.594	(0.011)	-0.137	(0.023)	0.000	(0.000)	1.679	(0.038)
									-1.679	(0.032)

Table B-1 (continued)
IRT Parameters for the Trial State Assessment
Reading for Literary Experience Scale, Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R015804	RI	4	0.571	(0.011)	0.705	(0.020)	0.000	(0.000)	2.414	(0.040)
									-0.392	(0.032)
									-2.022	(0.072)
R015805	RI	5	0.965	(0.058)	0.242	(0.050)	0.288	(0.020)		
R015806	RI	6	0.617	(0.013)	0.358	(0.022)	0.000	(0.000)	1.377	(0.033)
									-1.377	(0.036)
R015807	RI	7	0.585	(0.015)	-0.162	(0.023)	0.000	(0.000)	1.170	(0.039)
									-1.170	(0.033)
R015808	RI	8	0.610	(0.035)	-1.662	(0.145)	0.219	(0.046)		
R015809	RI	9	0.580	(0.014)	0.017	(0.026)	0.000	(0.000)	1.435	(0.042)
									-1.435	(0.038)

Table B-2
IRT Parameters for the Trial State Assessment
Reading to Gain Information Scale, Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R012201	RF	1	0.206	(0.014)	-1.355	(0.115)	0.000	(0.000)		
R012202	RF	2	0.819	(0.049)	0.440	(0.051)	0.213	(0.019)		
R012203	RF	3	0.800	(0.050)	0.594	(0.048)	0.193	(0.018)		
R012204	RF	4	0.424	(0.011)	-0.011	(0.019)	0.000	(0.000)	1.356	(0.050)
									-0.465	(0.044)
									-0.891	(0.050)
R012205	RF	5	1.296	(0.072)	0.563	(0.032)	0.277	(0.013)		
R012206	RF	6	1.102	(0.031)	0.729	(0.019)	0.000	(0.000)		
R012207	RF	7	0.529	(0.034)	-0.755	(0.148)	0.237	(0.043)		
R012208	RF	8	0.868	(0.025)	-0.453	(0.024)	0.000	(0.000)		
R012209	RF	9	1.187	(0.058)	0.336	(0.034)	0.178	(0.015)		
R012210	RF	10	0.601	(0.024)	-1.550	(0.060)	0.000	(0.000)		
R012501	RJ	1	0.925	(0.224)	2.799	(0.346)	0.310	(0.008)		
R012502	RJ	2	0.850	(0.041)	-2.035	(0.107)	0.234	(0.047)		
R012503	RJ	3	0.982	(0.025)	-0.085	(0.018)	0.000	(0.000)		
R012504	RJ	4	0.663	(0.020)	-0.373	(0.026)	0.000	(0.000)		
R012505	RJ	5	1.125	(0.050)	-0.840	(0.055)	0.256	(0.026)		
R012506	RJ	6	0.765	(0.022)	-0.220	(0.023)	0.000	(0.000)		
R012507	RJ	7	1.213	(0.060)	-0.453	(0.052)	0.357	(0.023)		
R012508	RJ	8	0.979	(0.026)	-0.469	(0.021)	0.000	(0.000)		
R012509	RJ	9	0.758	(0.045)	-0.617	(0.103)	0.322	(0.035)		
R012510	RJ	10	0.929	(0.050)	-0.435	(0.071)	0.306	(0.028)		
R012511	RJ	11	0.940	(0.028)	-0.682	(0.025)	0.000	(0.000)		
R012512	RJ	12	0.363	(0.011)	0.563	(0.024)	0.000	(0.000)	0.861	(0.057)
									0.226	(0.057)
									-1.088	(0.070)
R012701	RG	1	1.167	(0.056)	0.002	(0.041)	0.286	(0.018)		
R012702	RG	2	0.590	(0.020)	-1.273	(0.044)	0.000	(0.000)		
R012703	RG	3	1.093	(0.029)	0.629	(0.018)	0.000	(0.000)		
R012704	RG	4	1.325	(0.063)	0.729	(0.023)	0.144	(0.010)		
R012705	RG	5	1.347	(0.045)	1.272	(0.023)	0.000	(0.000)		
R012706	RG	6	0.642	(0.025)	1.335	(0.045)	0.000	(0.000)		
R012707	RG	7	2.225	(0.102)	0.396	(0.019)	0.245	(0.010)		
R012708	RG	8	0.673	(0.018)	1.634	(0.021)	0.000	(0.000)	1.253	(0.028)
									0.386	(0.036)
									-1.639	(0.108)
R012709	RG	9	0.548	(0.049)	0.052	(0.147)	0.288	(0.040)		
R012710	RG	10	1.075	(0.037)	0.839	(0.023)	0.000	(0.000)		
R015701	RH	1	1.030	(0.059)	-0.674	(0.079)	0.458	(0.028)		
R015702	RH	2	0.575	(0.011)	-0.040	(0.023)	0.000	(0.000)	1.638	(0.036)
									-1.638	(0.033)
R015703	RH	3	0.680	(0.012)	0.023	(0.021)	0.000	(0.000)	1.651	(0.032)
									-1.651	(0.030)

Table B-2 (continued)
IRT Parameters for the Trial State Assessment
Reading to Gain Information Scale, Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R015704	RH	4	0.617	(0.015)	-0.214	(0.018)	0.000	(0.000)	0.394	(0.032)
									-0.394	(0.028)
R015705	RH	5	0.730	(0.017)	0.154	(0.017)	0.000	(0.000)	0.808	(0.027)
									-0.808	(0.026)
R015706	RH	6	0.921	(0.066)	1.099	(0.039)	0.192	(0.013)		
R015707	RH	7	0.511	(0.012)	0.246	(0.024)	0.000	(0.000)	1.235	(0.037)
									-1.235	(0.039)
R015708	RH	8	0.587	(0.032)	-0.194	(0.084)	0.147	(0.028)		
R015709	RH	9	0.439	(0.016)	1.089	(0.035)	0.000	(0.000)	0.400	(0.043)
									-0.400	(0.057)

Table B-3
IRT Parameters for the National Reading Samples
Reading for Literary Experience Scale, Age 9/Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R012001	RC	1	1.786	(0.087)	0.683	(0.018)	0.102	(0.009)		
R012002	RC	2	1.668	(0.050)	0.071	(0.014)	0.000	(0.000)		
R012003	RC	3	1.871	(0.085)	-0.348	(0.029)	0.201	(0.017)		
R012004	RC	4	0.822	(0.031)	0.564	(0.026)	0.000	(0.000)		
R012005	RC	5	1.325	(0.072)	0.292	(0.032)	0.208	(0.015)		
R012006	RC	6	0.575	(0.018)	0.941	(0.023)	0.000	(0.000)	0.175	(0.041)
									0.058	(0.052)
									-0.233	(0.063)
R012007	RC	7	0.883	(0.056)	-0.341	(0.074)	0.251	(0.028)		
R012008a*	RC	8	0.634	(0.035)	-0.295	(0.050)	0.000	(0.000)		
R012008b*	RC	8	1.067	(0.053)	-0.393	(0.037)	0.000	(0.000)		
R012009	RC	9	1.589	(0.095)	-0.458	(0.050)	0.335	(0.024)		
R012010a*	RC	10	1.403	(0.066)	-0.045	(0.028)	0.000	(0.000)		
R012010b*	RC	10	1.281	(0.064)	-0.348	(0.033)	0.000	(0.000)		
R012011	RC	11	2.018	(0.117)	-0.041	(0.031)	0.256	(0.018)		
R012101	RD	1	2.073	(0.111)	-0.667	(0.035)	0.319	(0.021)		
R012102a*	RD	2	0.941	(0.044)	0.167	(0.030)	0.000	(0.000)		
R012102b*	RD	2	0.894	(0.042)	-0.928	(0.048)	0.000	(0.000)		
R012103	RD	3	1.445	(0.066)	-0.284	(0.035)	0.195	(0.018)		
R012104	RD	4	0.795	(0.028)	-0.107	(0.026)	0.000	(0.000)		
R012105	RD	5	0.898	(0.054)	0.126	(0.052)	0.180	(0.021)		
R012106a*	RD	6	1.140	(0.052)	0.222	(0.027)	0.000	(0.000)		
R012106b*	RD	6	0.950	(0.049)	0.523	(0.032)	0.000	(0.000)		
R012107	RD	7	1.298	(0.079)	0.316	(0.037)	0.245	(0.017)		
R012108	RD	8	0.720	(0.027)	-0.925	(0.043)	0.000	(0.000)		
R012109	RD	9	0.649	(0.026)	-0.892	(0.047)	0.000	(0.000)		
R012110	RD	10	0.899	(0.059)	-0.824	(0.099)	0.302	(0.036)		
R012111	RD	11	1.083	(0.037)	1.455	(0.023)	0.000	(0.000)	0.843	(0.022)
									-0.843	(0.060)
R012112	RD	12	0.869	(0.037)	-0.473	(0.037)	0.000	(0.000)		
R012401b**	RI	1	1.027	(0.044)	1.861	(0.031)	0.000	(0.000)	1.267	(0.029)
									-0.421	(0.079)
									-0.845	(0.303)
R012402b**	RI	2	1.019	(0.101)	0.071	(0.087)	0.333	(0.033)		
R012403b**	RI	3	1.221	(0.066)	0.940	(0.034)	0.000	(0.000)		
R012404b**	RI	4	1.164	(0.097)	0.270	(0.055)	0.217	(0.025)		
R012405b**	RI	5	1.362	(0.144)	0.900	(0.046)	0.211	(0.019)		
R012406b**	RI	6	0.977	(0.051)	0.460	(0.032)	0.000	(0.000)		
R012407b**	RI	7	1.071	(0.052)	-0.188	(0.033)	0.000	(0.000)		
R012408b**	RI	8	1.608	(0.145)	0.303	(0.048)	0.287	(0.024)		
R012409b**	RI	9	1.462	(0.080)	0.702	(0.028)	0.000	(0.000)		
R012601	RE	1	0.905	(0.039)	1.236	(0.039)	0.000	(0.000)		

***Note:** a = item parameters are based on only 1994 data; b = item parameters are based on only 1992 data.*
****Note:** This block name identifies two different sets of items for 1992 and 1994—a = 1994 items; b = 1992 items.*

Table B-3 (continued)
IRT Parameters for the National Reading Samples
Reading for Literary Experience Scale, Age 9/Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R012602	RE	2	1.931	(0.099)	1.341	(0.031)	0.193	(0.007)		
R012603	RE	3	1.623	(0.085)	0.269	(0.027)	0.211	(0.015)		
R012604	RE	4	1.197	(0.050)	1.196	(0.030)	0.000	(0.000)		
R012605	RE	5	1.220	(0.116)	1.038	(0.043)	0.312	(0.015)		
R012606	RE	6	1.520	(0.095)	0.476	(0.031)	0.270	(0.015)		
R012607	RE	7	1.135	(0.036)	1.394	(0.018)	0.000	(0.000)	0.639	(0.022)
									0.201	(0.031)
									-0.840	(0.074)
R012608	RE	8	0.900	(0.071)	-0.177	(0.091)	0.373	(0.031)		
R012609	RE	9	1.055	(0.092)	0.851	(0.045)	0.230	(0.018)		
R012610	RE	10	2.405	(0.145)	0.747	(0.025)	0.395	(0.012)		
R012611a*	RE	11	0.874	(0.050)	0.408	(0.038)	0.000	(0.000)		
R012611b*	RE	11	0.855	(0.049)	-0.129	(0.041)	0.000	(0.000)		
R015801a**	RI	1	1.109	(0.077)	-1.040	(0.086)	0.228	(0.035)		
R015802a**	RI	2	0.563	(0.031)	-0.556	(0.056)	0.000	(0.000)		
R015803a**	RI	3	0.645	(0.021)	0.064	(0.035)	0.000	(0.000)	1.438	(0.056)
									-1.438	(0.052)
R015804a**	RI	4	0.731	(0.024)	0.919	(0.029)	0.000	(0.000)	1.948	(0.050)
									-0.255	(0.046)
									-1.693	(0.116)
R015805a**	RI	5	1.055	(0.111)	0.472	(0.068)	0.317	(0.026)		
R015806a**	RI	6	0.681	(0.026)	0.452	(0.035)	0.000	(0.000)	1.132	(0.050)
									-1.132	(0.057)
R015807a**	RI	7	0.680	(0.029)	0.077	(0.034)	0.000	(0.000)	0.992	(0.056)
									-0.992	(0.051)
R015808a**	RI	8	0.843	(0.068)	-0.938	(0.117)	0.218	(0.040)		
R015809a**	RI	9	0.639	(0.026)	0.247	(0.042)	0.000	(0.000)	1.334	(0.064)
									-1.334	(0.063)

***Note:** a = item parameters are based on only 1994 data; b = item parameters are based on only 1992 data.*
****Note:** This block name identifies two different sets of items for 1992 and 1994—a = 1994 items; b = 1992 items.*

Table B-4

IRT Parameters for the National Reading Samples
Reading to Gain Information Scale, Age 9/Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R012201a*	RF	1	0.272	(0.024)	-0.484	(0.100)	0.000	(0.000)		
R012201b*	RF	1	0.399	(0.029)	0.142	(0.064)	0.000	(0.000)		
R012202	RF	2	0.819	(0.060)	0.534	(0.059)	0.224	(0.021)		
R012203	RF	3	0.797	(0.061)	0.682	(0.056)	0.200	(0.020)		
R012204	RF	4	0.523	(0.015)	0.385	(0.021)	0.000	(0.000)	1.114	(0.046)
									-0.256	(0.045)
									-0.858	(0.057)
R012205	RF	5	1.331	(0.090)	0.647	(0.036)	0.264	(0.015)		
R012206	RF	6	1.309	(0.046)	0.800	(0.021)	0.000	(0.000)		
R012207	RF	7	0.573	(0.043)	-0.530	(0.137)	0.222	(0.041)		
R012208	RF	8	0.906	(0.033)	-0.197	(0.027)	0.000	(0.000)		
R012209a*	RF	9	1.412	(0.106)	0.475	(0.041)	0.146	(0.018)		
R012209b*	RF	9	1.724	(0.162)	0.515	(0.044)	0.264	(0.021)		
R012210	RF	10	0.767	(0.032)	-0.965	(0.047)	0.000	(0.000)		
R012301b**	RH	1	0.805	(0.079)	-0.020	(0.112)	0.293	(0.039)		
R012302b**	RH	2	1.052	(0.074)	-0.325	(0.069)	0.204	(0.032)		
R012303b**	RH	3	1.170	(0.051)	-0.260	(0.028)	0.000	(0.000)		
R012304b**	RH	4	1.716	(0.323)	2.188	(0.200)	0.244	(0.010)		
R012305b**	RH	5	0.527	(0.018)	1.131	(0.036)	0.000	(0.000)	2.450	(0.066)
									0.130	(0.058)
									-2.580	(0.176)
R012306b**	RH	6	0.834	(0.047)	0.940	(0.046)	0.000	(0.000)		
R012307b**	RH	7	1.384	(0.093)	-0.001	(0.046)	0.172	(0.024)		
R012308b**	RH	8	0.913	(0.048)	0.540	(0.035)	0.000	(0.000)		
R012309b**	RH	9	0.920	(0.123)	0.985	(0.078)	0.249	(0.027)		
R012310b**	RH	10	0.916	(0.054)	0.620	(0.039)	0.000	(0.000)		
R012501	RJ	1	0.539	(0.172)	3.749	(0.773)	0.285	(0.014)		
R012502	RJ	2	1.153	(0.061)	-1.382	(0.073)	0.218	(0.034)		
R012503a*	RJ	3	1.219	(0.053)	0.177	(0.025)	0.000	(0.000)		
R012503b*	RJ	3	1.114	(0.052)	0.626	(0.030)	0.000	(0.000)		
R012504	RJ	4	0.861	(0.029)	0.042	(0.023)	0.000	(0.000)		
R012505	RJ	5	1.274	(0.063)	-0.559	(0.047)	0.222	(0.023)		
R012506	RJ	6	0.928	(0.031)	0.077	(0.022)	0.000	(0.000)		
R012507	RJ	7	1.322	(0.072)	-0.304	(0.047)	0.294	(0.022)		
R012508	RJ	8	1.125	(0.037)	-0.135	(0.021)	0.000	(0.000)		
R012509	RJ	9	0.631	(0.045)	-0.727	(0.134)	0.240	(0.042)		
R012510	RJ	10	0.962	(0.062)	-0.264	(0.072)	0.303	(0.028)		
R012511	RJ	11	1.129	(0.040)	-0.326	(0.024)	0.000	(0.000)		
R012512a*	RJ	12	0.409	(0.021)	0.687	(0.039)	0.000	(0.000)	0.910	(0.088)
									0.195	(0.088)
									-1.106	(0.115)

***Note:** a = item parameters are based on only 1994 data; b = item parameters are based on only 1992 data.*

****Note:** This block name identifies two different sets of items for 1992 and 1994—a = 1994 items; b = 1992 items.*

Table B-4 (continued)
IRT Parameters for the National Reading Samples
Reading to Gain Information Scale, Age 9/Grade 4

NAEP ID	Block	Item	A	S.E.	B	S.E.	C	S.E.	D	S.E.
R012512b*	RJ	12	0.451	(0.024)	1.140	(0.044)	0.000	(0.000)	0.842	(0.076)
									0.390	(0.085)
									-1.232	(0.136)
R012701	RG	1	1.351	(0.072)	0.107	(0.037)	0.262	(0.018)		
R012702	RG	2	0.635	(0.024)	-0.950	(0.044)	0.000	(0.000)		
R012703	RG	3	1.168	(0.040)	0.791	(0.022)	0.000	(0.000)		
R012704	RG	4	1.716	(0.099)	0.913	(0.024)	0.169	(0.010)		
R012705a*	RG	5	1.608	(0.094)	1.330	(0.036)	0.000	(0.000)		
R012705b*	RG	5	1.283	(0.102)	1.820	(0.080)	0.000	(0.000)		
R012706a*	RG	6	0.648	(0.046)	1.543	(0.089)	0.000	(0.000)		
R012706b*	RG	6	0.828	(0.048)	0.957	(0.047)	0.000	(0.000)		
R012707	RG	7	2.473	(0.147)	0.559	(0.022)	0.267	(0.012)		
R012708a*	RG	8	0.742	(0.035)	1.852	(0.037)	0.000	(0.000)	1.250	(0.044)
									0.382	(0.063)
									-1.633	(0.231)
R012708b*	RG	8	1.050	(0.048)	1.889	(0.031)	0.000	(0.000)	1.343	(0.032)
									-0.118	(0.065)
									-1.226	(0.297)
R012709	RG	9	0.591	(0.053)	0.034	(0.127)	0.229	(0.039)		
R012710a*	RG	10	1.064	(0.068)	1.004	(0.044)	0.000	(0.000)		
R012710b*	RG	10	1.141	(0.069)	0.779	(0.038)	0.000	(0.000)		
R015701a**	RH	1	1.000	(0.081)	-0.691	(0.102)	0.299	(0.039)		
R015702a**	RH	2	0.697	(0.022)	0.229	(0.033)	0.000	(0.000)	1.405	(0.049)
									-1.405	(0.051)
R015703a**	RH	3	0.763	(0.025)	0.216	(0.031)	0.000	(0.000)	1.325	(0.046)
									-1.325	(0.047)
R015704a**	RH	4	0.728	(0.031)	0.108	(0.026)	0.000	(0.000)	0.326	(0.047)
									-0.326	(0.044)
R015705a**	RH	5	0.851	(0.036)	0.395	(0.026)	0.000	(0.000)	0.664	(0.040)
									-0.664	(0.042)
R015706a**	RH	6	1.337	(0.140)	1.242	(0.053)	0.230	(0.015)		
R015707a**	RH	7	0.619	(0.025)	0.519	(0.036)	0.000	(0.000)	1.112	(0.053)
									-1.112	(0.061)
R015708a**	RH	8	0.706	(0.065)	0.123	(0.096)	0.171	(0.033)		
R015709a**	RH	9	0.583	(0.036)	1.215	(0.054)	0.000	(0.000)	0.324	(0.060)
									-0.324	(0.085)

***Note:** a = item parameters are based on only 1994 data; b = item parameters are based on only 1992 data.*
****Note:** This block name identifies two different sets of items for 1992 and 1994—a = 1994 items; b = 1992 items.*

APPENDIX C

GLOSSARY OF TERMS

assessment session. The period of time during which a NAEP booklet is administered to one or more individuals.

background questionnaires. The instruments used to collect information about students' demographics and educational experiences.

bias. In statistics, the difference between the expected value of an estimator and the population parameter being estimated. If the average value of the estimator over all possible samples (the estimator's expected value) equals the parameter being estimated, the estimator is said to be **unbiased**; otherwise, the estimator is **biased**.

BIB (Balanced Incomplete Block) spiraling. A complex variant of multiple matrix sampling, in which items are administered in such a way that each pair of items is administered to a nationally representative sample of respondents.

block. A group of assessment items created by dividing the item pool for an age/grade into subsets. Used in the implementation of the BIB spiral sample design.

booklet. The assessment instrument created by combining blocks of assessment items.

calibrate. To estimate the parameters of a set of items from responses of a sample of examinees.

clustering. The process of forming sampling units as groups of other units.

codebook. A formatted printout of NAEP data for a particular sample of respondents.

coefficient of variation. The ratio of the standard deviation of an estimate to the value of the estimate.

common block. A group of background items included in the beginning of every assessment booklet.

conditional probability. Probability of an event, given the occurrence of another event.

conditioning variables. Demographic and other background variables characterizing a respondent. Used in construction of plausible values.

constructed-response item. A nonmultiple-choice item that requires some type of written or oral response.

degrees of freedom. [of a variance estimator] The number of independent pieces of information used to generate a variance estimate.

derived variables. Subgroup data that were not obtained directly from assessment responses, but through procedures of interpretation, classification, or calculation.

design effects. The ratio of the variance for the sample design to the variance for a simple random sample of the same size.

distractor. An incorrect response choice included in a multiple-choice item.

excluded student questionnaire. An instrument completed for every student who was sampled but excluded from the assessment.

excluded students. Sampled students determined by the school to be unable to participate because they have limited English language proficiency, are judged as being mildly mentally retarded (educable), or are functionally disabled.

expected value. The average of the sample estimates given by an estimator over all

possible samples. If the estimator is unbiased, then its expected value will equal the population value being estimated.

field test. A pretest of items to obtain information regarding clarity, difficulty levels, timing, feasibility, and special administrative situations; performed before revising and selecting items to be used in the assessment.

focused-BIB spiraling. A variation of BIB spiraling in which items are administered in such a way that each pair of items *within a subject area* is administered to a nationally representative sample of respondents.

foils. The correct and incorrect response choices included in a multiple-choice item.

group effect. The difference between the mean for a group and the mean for the nation.

imputation. Prediction of a missing value according to some procedure, using a mathematical model in combination with available information. See **plausible values**.

imputed race/ethnicity. The race or ethnicity of an assessed student, as derived from his or her responses to particular common background items. A NAEP **reporting subgroup**.

item response theory (IRT). Test analysis procedures that assume a mathematical model for the probability that a given examinee will respond correctly to a given exercise.

jackknife. A procedure to estimate standard errors of percentages and other statistics. Particularly suited to complex sample designs.

machine-readable catalog. Computer processing control information, IRT parameters, foil codes, and labels in a computer-readable format.

metropolitan statistical area (MSA). An area defined by the federal government for the purposes of presenting general-purpose statistics for metropolitan areas. Typically, an MSA contains a city with a population of at least 50,000 plus adjacent areas.

multistage sample design. Indicates more than one stage of sampling. An example of three-stage sampling: 1) sample of counties (primary sampling units or PSUs); 2) sample of schools within each sample county; 3) sample of students within each sample school.

multiple matrix sampling. Sampling plan in which different samples of respondents take different samples of items.

NAEP scales. The scales common across age/grade levels and assessment years used to report NAEP results.

nonresponse. The failure to obtain responses or measurements for all sample elements.

nonsampling error. A general term applying to all sources of error except sampling error. Includes errors from defects in the sampling frame, response or measurement error, and mistakes in processing the data.

objective. A desirable education goal agreed upon by scholars in the field, educators, and concerned laypersons, and established through the consensus approach.

observed race/ethnicity. Race or ethnicity of an assessed student as perceived by the exercise administrator.

oversampling. Deliberately sampling a portion of the population at a higher rate than the remainder of the population.

parental education. The level of education of the mother and father of an assessed student as derived from the student's response to two assessment items. A NAEP **reporting subgroup**.

percent correct. The percent of a target population that would answer a particular exercise correctly.

plausible values. Proficiency values drawn at random from a conditional distribution of a NAEP respondent, given his or her response to cognitive exercises and a specified subset of background variables (conditioning variables). The selection of a plausible value is a form of **imputation**.

poststratification. Classification and weighting to correspond to external values of selected sampling units by a set of strata definitions after the sample has been selected.

primary sampling unit (PSU). The basic geographic sampling unit for NAEP. Either a single county or a set of contiguous counties.

probability sample. A sample in which every element of the population has a known, nonzero probability of being selected.

pseudoreplicate. The value of a statistic based on an altered sample. Used by the **jackknife** variance estimator.

QED. Quality Education Data, Inc. A supplier of lists of schools, school districts, and other school data.

random variable. A variable that takes on any value of a specified set with a particular probability.

region. One of four geographic areas used in gathering and reporting data: Northeast, Southeast, Central, and West (as defined by the Office of Business Economics, U.S. Department of Commerce). A NAEP **reporting subgroup.**

reporting subgroup. Groups within the national population for which NAEP data are reported: for example, gender, race/ethnicity, grade, age, level of parental education, region, and type of location.

respondent. A person who is eligible for NAEP, is in the sample, and responds by completing one or more items in an assessment booklet.

response options. In a multiple-choice question, alternatives that can be selected by a respondent.

sample. A portion of a population, or a subset from a set of units, selected by some probability mechanism for the purpose of investigating the properties of the population. NAEP does not assess an entire population but rather selects a representative sample from the group to answer assessment items.

sampling error. The error in survey estimates that occurs because only a sample of the population is observed. Measured by sampling **standard error**.

sampling frame. The list of sampling units from which the sample is selected.

sampling weight. A multiplicative factor equal to the reciprocal of the probability of a respondent being selected for assessment with adjustment for nonresponse and perhaps also for poststratification. The sum of the weights provides an estimate of the number of persons in the population represented by a respondent in the sample.

school characteristics and policy questionnaire. A questionnaire completed for each school by the principal or other official; used to gather information concerning school administration, staffing patterns, curriculum, and student services.

secondary-use data files. Computer files containing respondent-level cognitive, demographic, and background data. Available for use by researchers wishing to perform analyses of NAEP data.

selection probability. The chance that a particular sampling unit has of being selected in the sample.

session. A group of students reporting for the administration of an assessment. Most schools conducted only one session, but some large schools conducted as many as 10 or more.

simple random sample. Process for selecting n sampling units from a population of N sampling units so that each sampling unit has an equal chance of being in the sample and every combination of n sampling units has the same chance of being in the sample chosen.

standard error. A measure of sampling variability and measurement error for a statistic. Because of NAEP's complex sample design, sampling standard errors are estimated by **jackknifing** the samples from first-stage sample estimates. Standard errors may also include a component due to the error of measurement of individual scores estimated using plausible values.

stratification. The division of a population into parts, called strata.

stratified sample. A sample selected from a population that has been stratified, with a sample selected independently in each stratum. The strata are defined for the purpose of reducing sampling error.

student ID number. A unique identification number assigned to each respondent to preserve his or her anonymity. NAEP does not record the names of any respondents.

subject area. One of the areas assessed by National Assessment; for example, art, civics, computer competence, geography, literature, mathematics, music, reading, science, U.S. history, or writing.

systematic sample (systematic random sample). A sample selected by a systematic method; for example, when units are selected from a list at equally spaced intervals.

teacher questionnaire. A questionnaire completed by selected teachers of sample students; used to gather information concerning years of teaching experience, frequency of assignments, teaching materials used, and availability and use of computers.

Trial State Assessment Program. The NAEP program, authorized by Congress in 1988, which was established to provide for a program of voluntary state-by-state assessments on a trial basis.

trimming. A process by which extreme weights are reduced (trimmed) to diminish the effect of extreme values on estimates and estimated variances.

type of location (TOL). One of the NAEP **reporting subgroups**, dividing the communities in the nation into groups on the basis of the proportion of the students living in each of three sizes and types of communities.

variance. The average of the squared deviations of a random variable from the expected value of the variable. The variance of an estimate is the squared standard error of the estimate.

APPENDIX D

UNWEIGHTED NONPUBLIC-SCHOOL DATA FILES

Introduction

Minimum sample size requirements for reporting nonpublic-school data were not met by the data from six jurisdictions that participated in the 1994 Trial State Assessment. For these jurisdictions, nonpublic-school data are provided separately from the public-school data. This appendix describes the data files that contain the nonpublic-school data for those jurisdictions and warns against analyzing these data or drawing conclusions from them.

Background Information

Minimum sample size requirements for reporting nonpublic-school data consist of two components: 1) a school sample size of six or more participating schools and 2) an assessed student sample size of at least 62.[7] The nonpublic-school data for six jurisdictions that participated in the 1994 Trial State Assessment (Arizona, New Hampshire, North Carolina, Tennessee, Texas, and Utah) included insufficient numbers of cooperating nonpublic schools. The data from North Carolina and Utah also did not contain sufficient numbers of assessed nonpublic school students. The participation rates for these nonpublic-school samples were too low to necessitate the creation of sampling weights, and proficiency estimates and replicate weights could not be calculated for them.

As a result, the raw data files for these six states contain only the public-school data, while the nonpublic school data is provided separately.

The nonpublic-school data are in separate files (and reside in a separate directory on the CD-ROM), so that they are automatically excluded from routine analyses. These data have very limited utility. They are included in the secondary-use data package for completeness because they were collected as part of the 1994 Trial State Assessment.

[7] Please see Appendix B of the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* (Mazzeo, Allen, & Kline, 1995) for more information about guidelines on the publication of NAEP results and 1994 Trial State Assessment participation rates.

Cautions

It is difficult to imagine an analysis in which these data could be used in a manner that is defensible or desirable. The goal of the Trial State design is to obtain a sample of students for each jurisdiction from which estimates of population and subpopulation characteristics can be obtained. Following the collection of assessment and background data from and about assessed and excluded students in a jurisdiction, sampling weights and associated sets of replicate weights are derived. ***The sampling weights are needed to make valid inferences from the student samples to the respective populations from which they were drawn and must be used for all analyses,*** whether exploratory or confirmatory. The replicate weights are used in the estimation of sampling variance.

Normally, each student is assigned a weight to be used for making inferences about the state's students. This weight is known as the *full-sample* or *overall* weight. In addition to estimation weights, a set of replicate weights would normally be provided for each student. These replicate weights are used in calculating the sampling errors of estimates obtained from the data, using the jackknife repeated replication method.

Because minimum sample size requirements were not met, the nonpublic-school data for these six states are *not* a representative sample of nonpublic-school fourth-grade students of the jurisdictions under consideration. They cannot be weighted to enable researchers to estimate parameters for those target populations.

Format of Data Files

The format and file naming conventions of the nonpublic-school data files for Arizona, New Hampshire, North Carolina, Tennessee, Texas, and Utah, and the contents of the accompanying file layout and data codebook for each data file are described in Chapter 9. The names and other characteristics of the files are given in Tables 9-1 and D-1. These data files reside in a separate directory (when provided on CD-ROM); in place of the FIPS code abbreviation that identifies the raw data files for each jurisdiction, the designations P1 - P6 are used to identify these files. Record lengths for these files are listed in Table 9-1. Note that the number of records for the raw data files varies by jurisdiction; Table D-1 lists the record count for each file, as well as the two-character designation for each state next to the state name.

Table D-1
Special Data File Record Counts by State

Special Abbreviation	State Name	Separate Nonpublic-School Grade 4 Data File Record Counts		
		Student	**Excluded Student**	**School**
P1	Arizona	91	3	3
P2	New Hampshire	116	0	5
P3	North Carolina	49	0	2
P4	Tennessee	83	0	4
P5	Texas	79	0	3
P6	Utah	32	0	1
Total		**450**	**3**	**18**

REFERENCES

CITED IN TEXT

Allen, N. L., Mazzeo J., Isham, S. P., Fong, Y. F., & Bowker, D. W. (1994). Data analysis and scaling for the 1992 trial state assessment in reading. In E. G. Johnson, J. Mazzeo, & D. Kline, *Technical report of the NAEP 1992 trial state assessment program in reading*. Washington, DC: National Center for Education Statistics..

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 26(2),* 163-175.

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15,* 9-38.

Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly.* (No. 17-TR-21) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

Johnson, E. G., Mislevy, R. J., & Zwick, R. (1990). Estimation of the standard errors of the adjusted 1986 NAEP results. In A. E. Beaton & R. Zwick, *Disentangling the NAEP 1985-86 reading anomaly.* (No. 19-TR-20). Princeton, NJ: Educational Testing Service.

Keyfitz, N. (1951). Sampling with probability proportional to size; adjustment for changes in probabilities. *Journal of the American Statistical Association, 46,* 105-109.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Kish, L. & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B, 36,* 1-22.

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37,* 218-220.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mazzeo, J. (1991). Data analysis and scaling. In S. L. Koffler, *The technical report of NAEP's 1990 Trial State Assessment program* (No. ST-21-01). Washington, DC: National Center for Education Statistics.

Mazzeo, J., Allen, N. L., & Kline, D. L. (1995). *Technical report of the NAEP 1994 Trial State Assessment Program in reading.* Washington, DC: National Center for Education Statistics.

Mazzeo, J., Chang, H., Kulick, E., Fong, Y. F., & Grima, A. (1993). Data analysis and scaling for the 1992 Trial State Assessment in mathematics. In E. G. Johnson, J. Mazzeo, & D. L. Kline, *Technical report of the NAEP 1992 Trial State Assessment program in mathematics.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mazzeo, J., Johnson, E. G., Bowker, D., & Fong, Y. F. (1992). *The use of collateral information in proficiency estimation for the Trial State Assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177-196.

Mislevy, R. J. (1990). Scaling procedures. In E.G. Johnson and R. Zwick, *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993-997.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17(2),* 131-154.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A.E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (No. 15-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J., & Wu, P-K. (1988). *Inferring examinee ability when some item responses are missing* (ETS Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2),* 159-176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data*. Chicago, IL: Scientific Software, Inc.

Rogers, A. M. (1991). *NAEP-MGROUP: Enhanced version of Sheehan's software for the estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6,* 309-316.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program] Princeton, NJ: Educational Testing Service.

Sheehan, K. M., & Mislevy, R. J. (1988). *Some consequences of the uncertainty in IRT linking procedures*. (ETS Research Report No. RR-88-38-ONR). Princeton, NJ: Educational Testing Service.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys.* New York: John Wiley & Sons.

Thomas, N. (1992). *Higher order asymptotic corrections applied in an EM algorithm for estimating educational proficiencies*. Unpublished manuscript.

Tsutakawa, R. K., & Johnson, J. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55,* 371-390.

Wingersky, M., Kaplan, B. A., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report*. (No 15-TR-20) Princeton, NJ: National Association of Educational Progress, Educational Testing Service.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17(2),* 155-173.